

DOCUMENT RESUME

ED 379 346

TM 022 740

AUTHOR Rudner, Lawrence M.; And Others
TITLE Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results. Research and Development Report.
INSTITUTION LMP Associates, Inc., Chevy Chase, MD.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO ISBN-0-16-045446-8; NCES-95-713
PUB DATE Jan 95
CONTRACT R999B20006
NOTE 112p.
AVAILABLE FROM U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)
EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Difficulty Level; *Educational Assessment; *Goodness of Fit; *Measurement Techniques; Models; National Surveys; *Responses; Statistical Analysis; *Test Items
IDENTIFIERS Accuracy; *National Assessment of Educational Progress; *Person Fit Measures; Weighting (Statistical)

ABSTRACT

Fit statistics provide a direct measure of assessment accuracy by analyzing the fit of measurement models to an individual's (or group's) response pattern. Students that lose interest during the assessment, for example, will miss exercises that are within their abilities. Such students will respond correctly to some more difficult items and incorrectly to some less difficult items. Most assessment programs, including the National Assessment of Educational Progress (NAEP), currently either ignore such response anomalies or assume they do not exist. The use of a weighted-total-fit-mean-square as a measure of assessment accuracy was investigated using data from the 1990 and 1992 NAEP assessments. The distribution of fit across individuals was examined for fit and item-type differences, and the practical significance of this type of fit statistic was explored. It is concluded that this person-fit statistic has little to offer in the analysis of traditional NAEP data. Sixteen tables present analysis results. Appendix A contains 12 subscale tables, and Appendix B presents software routines. (Contains 62 references.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

BEST COPY AVAILABLE

NATIONAL CENTER FOR EDUCATION STATISTICS

Research and Development Report

January 1995

Use of Person-Fit Statistics in Reporting and Analyzing National Assessment of Educational Progress Results

Lawrence M. Rudner
LMP Associates and Catholic University of America

Gary Skagg
West Mesa Associates

Gerald Bracey
Consultant

Pamela R. Getson
Children's Hospital and National Medical Center

**U.S. Department of Education
Office of Educational Research and Improvement**

NCES 95-713

U.S. Department of Education

Richard W. Riley

Secretary

Office of Educational Research and Improvement

Sharon P. Robinson

Assistant Secretary

National Center for Education Statistics

Emerson J. Elliott

Commissioner

National Center for Education Statistics

"The purpose of the Center shall be to collect, analyze, and disseminate statistics and other data related to education in the United States and in other nations."—Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).

January 1995

Contact:

Alex Sedlacek

(202) 219-1734

This research was supported with funds from the Office of Educational Research and Improvement, U.S. Department of Education, Grant No. R999B20006. The views and opinions expressed in this paper are those of the first author and do not necessarily reflect the views of the other authors or the funding agency. Address correspondence to Lawrence Rudner, LMP Associates, 3109 Rolling Road #201, Chevy Chase, MD 20815.

Foreword

The Research and Development (R&D) series of reports has been initiated

- 1) To share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available.
- 2) To share results that are, to some extent, on the "cutting edge" of methodological developments. Emerging analytical approaches and new computer software developments often permit new, and sometimes controversial analysis to be done. By participating in "frontier research," we hope to contribute to the resolution of issues and improved analysis.
- 3) To participate in discussions of emerging issues of interest to educational researchers, statisticians, and the Federal statistical community in general. Such reports may document workshops and symposiums sponsored by NCES that address methodological and analytical issues or may share and discuss issues regarding NCES practice, procedures, and standards.

The common theme in all these goals is that these reports present results or discussion that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore the techniques and inferences made from the data are tentative and are subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done. Such responses should be addressed to:

Emerson Elliott
Commissioner
National Center for Education Statistics
555 New Jersey Ave. NW
Washington, D.C. 20208

Abstract

Fit statistics provide a direct measure of assessment accuracy by analyzing the fit of measurement models to an individual's (or group's) response pattern. Students that lose interest during the assessment, for example, will miss exercises that are within their abilities. Such students will respond correctly to some more difficult items and incorrectly to some less difficult items. Most assessment programs, including NAEP, currently either ignore such response anomalies or assume they do not exist.

We investigated the use of a weighted-total-fit-mean-square as a measure of assessment accuracy using data from the 1990 and 1992 NAEP assessments. We examined the distribution of fit across individuals, looked for group and item-type differences, and investigated the practical significance of this type of fit statistic. We conclude that this person-fit statistic has little to offer in the analysis of traditional NAEP data.

Table of Contents

Foreword	iii
Abstract	iv
Introduction	1
Organization of this report	2
Historical Background	3
This project	6
Related Literature	8
What is person-fit?	8
Why might person-fit statistics be needed?	10
Selecting appropriate fit statistics	12
Is fit a transitory or stable characteristic?	14
How has person-fit statistics been applied?	21
Observations	23
The research has been unsystematic.	23
The research has been atheoretical.	24
The research has not explored practical applications.	25
What fit statistics have been proposed?	25
Rasch Model Approaches	25
Birnbaum Model Approaches	26
Correlation Approaches	27
Sequence Approaches	27
Discussion	28
What are the implications for NAEP?	30
1990 Trial State Assessment	35
Methods	35
Data	35
Computation of fit	36
Results	36
Description of Fit Statistics	37
Fit by State	39
Fit by key demographic variables	41
Fit by Item Block Order	42
Block M5 and the Data Analysis subscale	43
Relationship of Person Fit to Proficiency	44
Prediction of Person Misfit	46
Effect of Trimming Misfitting Students on Population Proficiency	48
Full 1990 & 1992 Assessments	52

Methods	52
Data	52
Computation of fit	53
Results	53
Overall Fit	53
Rater versus no Rater	54
Constructed Response versus multiple choice questions	55
Calculator Use versus no Calculator	56
Correlations of fit statistics with total scaled score.	57
Item Fit	59
Do misfitting students alter the results?	60
Discussion	63
Appendix 1: Subscale Tables	66
Appendix 2: Software Routines	93
References	98

Tables in the Report

Table 1	Descriptive Statistics of Fit Mean Square Statistics	38
Table 2	Average mean square fit (and standard deviation) on the NAEP TSA by state	40
Table 3	Means (and Standard Deviations) of Fit Statistics by Select Demographic Characteristics	41
Table 4	Mean Fit Statistics of Booklets According to Block Order and Data Analysis Subscale	43
Table 5	Correlations Between Fit Statistics and Plausible Values	44
Table 6	Percent of poorly fitting students on each scale that are also poor fit on the other scales	46
Table 7	The best predictors of poor overall fit and the percent of respondents with poor fit by item response.	49
Table 8	Mean Subscale and Composite Proficiencies and Anchor Level Results for NAEP and Untrimmed and Trimmed Samples	50
Table 9	Trimmed and Untrimmed State Composite Scaled Means (and Standard Deviations)	51
Table 10	Distribution of the mean square fit statistic for five National Assessments	54
Table 11	Mean fit on rater versus no rater items	55
Table 12	Mean fit on constructed response versus multiple choice items	56
Table 13	Mean Fit on Items with and without Calculator Use	57
Table 14	Correlations (N) of various fit statistics and Scaled Score	58
Table 15	Mean Fit (and standard deviation) by item type	60
Table 16	NAEP Scales Score Mean Differences for untrimmed versus trimmed groups for 5 National Eighth Grade Assessments.	62
Table A-1	Correlations of Model Fit Statistics Between Subscales and Item Blocks on the 1990 Trial State Assessment	67
Table A-2	Correlations Between Fit Statistics and Plausible Values on the 1990 Trial State Assessment	67
Table A-3	Correlations Greater than .10 or Less than -.10 Between Subscale Fit Statistics and Background Variables on the 1990 Trial State Assessment	68
Table A-4	Results of Discriminant Analyses by Subscale on the 1990 Trial State Assessment	69
Table A-5	Means and Standard Deviations of Fit Statistics by Sex, Race and Subscale on the 1990 Trial State Assessment	72
Table A-6	Means and Standard Deviations of Fit Statistics by Select Demographic variables and Subscale on the 1990 Trial State Assessment	75

Table A-7	Mean Composite Proficiency and Anchor Level Results for NAEP and Untrimmed and Trimmed Samples on the 1990 Trial State Assessment	78
Table A-8	Trimmed and Untrimmed Mean scores for the 1990 NAEP Mathematics Assessment	83
Table A-9	Trimmed and Untrimmed Mean scores for the 1990 NAEP Reading Assessment	85
Table A-10	Trimmed and Untrimmed Mean scores for the 1992 NAEP Estimation Assessment	87
Table A-11	Trimmed and Untrimmed Mean scores for the 1992 NAEP Mathematics Assessment	89
Table A-12	Trimmed and Untrimmed Mean scores for the 1992 NAEP Reading Assessment	91

The use of person fit statistics in the analysis and reporting of NAEP results¹

Lawrence M. Rudner, LMP Associates and Catholic University of America

Gary Skaggs, West Mesa Associates

Gerald Bracey, Consultant

Pamela R. Getson, Children's Hospital and National Medical Center

Chapter 1: Introduction

The growing body of research on the use of fit statistics has significant implications for the National Assessment of Educational Progress. This family of statistics provides a direct measure of assessment accuracy by analyzing the fit of measurement models to an individual's (or group's) response pattern. Students that lose interest during the assessment, for example, will miss exercises that are within their abilities. Such students will respond correctly to some more difficult items and incorrectly to some less difficult items. Most assessment programs, including NAEP, currently either ignore such response anomalies or assume they do not exist.

We investigated the use of a weighted-total-fit-mean-square as a measure of assessment accuracy using data from the 1990 NAEP Trail State Assessment (TSA) and the full 1990 and 1992 Assessments. We examined the distribution of fit across individuals, looked for

This research was supported with funds from the Office of Educational Research and Improvement, U.S. Department of Education, Grant No. R999B20006. The views and opinions expressed in this paper are those of the first author and do not necessarily reflect the views of the other authors or the funding agency. Address correspondence to Lawrence Rudner, LMP Associates, 3109 Rolling Road #201, Chevy Chase, MD 20815.

The authors are indebted to the NAEP staff at the Educational Testing Service for the preparation of excellent datasets and documentation, Alex Sedlacek of OERI for her guidance and assistance with contracts management, and Alfred Rogers of the Educational Testing Service for his assistance in preparing pre-release datasets for analysis.

group and item-type differences, and investigated the practical significance of this type of fit statistic. Specifically, we sought to answer the following questions:

1. What is the distribution of the fit statistic for the 1990 and 1992 NAEP assessments?
2. What demographic, background, and attitude variables are most associated with poor fit?
3. Are there meaningful differences in the means and variances of assessment accuracy when grouped by key demographic, background, and attitude variables?
4. Are there significant differences in NAEP results after trimming the data of poorly assessed individuals, i.e. are noted group differences educationally meaningful?

Responses to these questions permit us to make recommendations as to whether fit statistics should be considered in analyzing or reporting NAEP results for traditional NAEP items. If, for example, there were large differences in the average fit across states, then an analysis and reporting that took fit into consideration might be warranted.

In this paper, we show that the fit of respondents to NAEP data is extremely good and that the fit statistic we analyzed had little to offer either in terms of analysis or in terms of reporting. Since there is a high correlation across person-fit statistics, we suspect that fit statistics in general have little to offer in the analysis or reporting of traditional NAEP items.

Organization of this report

The remainder of the introduction provides an historical background of the NAEP project and an overview of person fit statistics. We then provide an indepth review of the literature on person-fit statistics which includes applications, methods, an analysis, and potential for NAEP. We then do an indepth analysis of the use of the weighted fit mean square with the 1990 NAEP trial state assessment. The data indicate that person fit statistics have little to offer with this traditional multiple choice test. We then replicate our analysis using data from the full 1990 mathematics, 1990 reading, 1992 estimation, 1992 mathematics, and 1992

reading assessments. We again note that individuals tend to respond constantly and differences in aggregated scores were minor. We also noted that, in aggregate, individuals are equally consistent on items that involve a rater and items that do not, and on constructed response versus multiple choice questions. We note that there are major differences between items that involve calculator use and items that do not. We speculate that these differences are due to the quality of the item parameters rather than individual inconsistencies. Appendices contain further analysis by subtests and some of the software used to reorganize the data.

Historical Background

In September 1963, when the Commissioner of Education and the Carnegie Corporation of New York took the first step toward providing "a means of ascertaining the educational level" of young Americans, an assessment carried out by the federal government raised the specter of national standards. At the time, this was an anathema. Such an assessment based on output would go far beyond the traditional measures of quality, which had focused on input (e.g., per-student expenditures) and process (e.g., pupil-teacher ratio). The overwhelming reality was that the odds were against those who championed the idea. Their first priority had to be making a national assessment happen, not the precise shape "it" should take. Implementing a national assessment of almost any shape would be (and was) itself a heroic achievement.

In deference to the political climate of the times, the founders of NAEP built numerous safeguards into the assessment. NAEP was a bold experiment - an assessment rather than a test. Items, called "exercises," would carry the weight of the assessment. There would be no total scores; the individual items would be the focus of reporting. This focus on the rows (items) rather than the columns (persons) of the data matrix was revolutionary; item p-values were to be taken seriously as conveyers of information - it was to be of intrinsic interest that "X percent of 9-year-olds can do such-and-such". This was true criterion-referenced

measurement as originally discussed by Glasser (1963), in contrast to a norm-referenced assessment that reported individuals' or groups' performance in terms of grade equivalents, percentiles, etc. It followed that the scope and quality of NAEP exercises would be of paramount importance, and so a very sophisticated (for its time) consensual process was developed to generate objectives and create exercises.

In the late 1960's, the states were very wary of interstate comparisons. To help assure broad cooperation in and support for the new National Assessment, it was designed to preclude interstate comparisons. The smallest units to be compared would be the four broad, geographic regions. These safeguards protected the fledgling NAEP and helped it to flourish and become a model for states and other countries.

In response to a 1983 grant announcement calling for new ideas for improving NAEP, Messick, Beaton, and Lord (1984) outlined their concept of a NAEP that would be responsive to the times, use the latest advances in measurement theory, and that would hold great promise for future analytic thought. As the new NAEP awardee in 1984, the Educational Testing Service began implementing these concepts. Some of the most technically significant changes made to NAEP by the ETS technical staff are:

- effect sampling efficiencies through balanced incomplete block (BIB) spiralling
- sampling by grade level as well as by age
- random sampling within schools
- item response theory scaling within and across age levels
- estimation of covariances among exercises and background questions
- reporting results using behavioral anchors
- introduction of methods for estimating missing values

Many of these changes laid the foundation for simple, yet far-reaching, improvements. Reporting results in terms of behavioral activities, for example, was revolutionary. Describing national performance on a seemingly simplistic scale made it possible for parents, policymakers, and other non-measurement experts to understand the results of our national assessment.

In addition to recent technical changes, recommendations for policy and procedural changes were recently made by the Study Group on the National Assessment of Student Achievement and the National Academy of Education. Many of these changes were incorporated in the far-reaching *National Assessment of Educational Progress Improvement Act* passed by the 100th Congress of the United States.

After considering 46 commissioned papers, the Study Group on the National Assessment of Student Achievement outlined seven major recommendations:

- (1) maintain NAEP's continuity;
- (2) assess the core curriculum;
- (3) focus on transitional grades (4, 8, and 12) and sample out-of-school 17-year-olds, adults and private school students;
- (4) create an independent Educational Assessment Council, with members to be appointed by the Secretary of Education;
- (5) provide for add-on and school district assessments;
- (6) assess and provide for add-on assessment of private school students; and
- (7) increase federal funding. (Alexander and James, 1987)

The Study Group strongly recommended that achievement data in several curricular areas be collected on each state and the District of Columbia and that state and local assessments be linked with NAEP. The 100th Congress called for a major overhaul of the National Assessment. The prime responsibility for the National Assessment was moved from a grantee to the National Center for Education Statistics. State trial assessments were mandated for 1990 and 1992.

Interest in NAEP is now at an all time high. Funding has been increased and states are participating in pilot studies of state-by-state comparisons using NAEP items. The structure has also changed radically. The government now exercises greater control over NAEP. The National Assessment Governing Board is significantly more independent and staffed by qualified measurement professionals. The high profile, heightened expectations, and new structure underscore the need for NAEP to represent the best America has to offer in terms of technical sophistication.

This project

We investigated the applicability of a state-of-the-art technique for analyzing assessment accuracy. At the individual level, we compare the predicted probability of a correct response given by the measurement model employed by NAEP to the observation whether the student correctly responded to the question. At the group level, we aggregate the predicted probabilities and the observed proportions. Rudner (1983) and Drasgow, Levine, and McLaughlin (1987) describe several statistics that have been used to evaluate correspondence in this context.

There are a variety of reasons why an individual or group test score may not be an accurate estimate of ability. Wright (1977) cited tendencies such as "guessing, cheating, sleeping, fumbling, plodding, and cultural bias" as causes for invalid item responses. Sleepers get bored with a test and do poorly on later items; fumblers do poorly in the beginning because of confusion with test format; plodders never get to later items on a test. Levine and Rubin (1979) also argued that there are occasions when a student is so unlike other examinees that the resulting test score cannot be regarded as an appropriate ability measure. They cited tendencies such as improperly attained items, answer sheet alignment errors, exceptional creativity, and poor test taking strategies as causes for invalid item responses. Levine and Drasgow (1982) provided some additional examples: a low ability examinee who copies half of the answers from a more able neighbor; examinees who are shown the answers to some items before the examination; high ability examinees with atypical schooling; examinees with low English fluency; and examinees who are conservative in their use of partial information.

All of NAEP's group statistics are based on aggregated individual statistics. While some of the errors introduced by sampling and data imputation are expected to average to zero when summed over groups, the type of measurement errors we are discussing have unknown distributional properties. Thus, there is the very real possibility that some groups of students are assessed more accurately by NAEP than other groups. If NAEP moves into more high

stakes testing and reports results by state, district or school, accuracy of assessment should have major implications concerning the use of NAEP results.

Chapter 2: Related Literature

In this section we briefly introduce what we mean by assessment accuracy, outline some approaches to estimating fit statistics, and discuss the implications of this research for some of the psychometric and reporting issues confronting the National Assessment.

The National Assessment builds heavily on the Item Response Theory measurement model. One basic feature of the model is that it provides a probability of a correct response to a given item by an individual student. The accuracy of the assessment at the individual level can be gauged by comparing the predicted probability to the observation whether the student correctly responded to the question. When the assessment is accurate, there will be a high correspondence between the two. Similarly, accuracy of assessment can be gauged at the group level by aggregating the predicted probabilities and the observed proportions.

What is person-fit?

"Whenever we measure anything, whether in the physical, the biological, or the social sciences, that measurement contains a certain amount of chance error....Two sets of measurements of the same features of the same individuals will never exactly duplicate each other....However, at the same time, repeated measurements of a series of objects or individuals will ordinarily show some consistency."

Robert L. Thorndike in *Educational Measurement* by Linquist, 1951.

Robert Thorndike's words capture several themes that resound strongly throughout the history of psychometrics in America:

- ▶ the concern over the accuracy of measurement;
- ▶ the depiction of physical measurement as being analogous to psychological measurement, thus reflecting psychology's obsession with emulating physics and establishing itself as a "true" science; and

- ▶ the notion that somehow, a "true" score exists that is always obscured to some degree by "error," thus making it the psychometrician's task to minimize such error.

Later in his chapter, Thorndike also mirrored another strong tradition in psychometrics — the tendency to view people as individual agents who are capable of acting independently of their environment. He went on to list twenty-three general and specific, lasting and temporary characteristics of the individual that could affect test scores between one testing and another, but he listed no characteristic of settings or contexts that might occasionally affect test performance. Of course, from the traditional psychometric view, to the extent that settings are standardized, variations should be minimal or non-existent. However, recent developments in cognitive psychology raise serious questions about this approach to measurement.

The classical psychometric tradition that Lindquist and Thorndike represented has been extended into a framework known as *generalizability theory* (Cronbach, 1972, Shavelson, Webb and Rowley, 1989). Generalizability theory advanced the field of psychometrics in that it can simultaneously estimate error from several sources. The theory can be used in criterion situations where decisions will be made and in traditional situations that examine individual differences.

In 1989, Shavelson et al. wrote that generalizability theory assumes "steady state" behavior, stating that "*Most measurement approaches, including CT [Classical Theory] and GT [Generalizability Theory], assume that behavior remains constant over observations.*" When behavior does vary, the theory presumes that variations arise from people being in manifestly different situations or in learning situations where change is expected over time.

Although most previous research and reporting about error in measurement dealt with whether **items** fit, in the last fifteen years, significant interest arose over a different view of the issue — whether **people** who respond to the items "fit." Although this concern focused primarily on students whose response patterns did not fit the typical response pattern, some

part of the research also examined the fit of the same student to different tests taken over time.

Attempts to find methods of systematically identifying students with different response patterns led to developing a number of *person-fit* statistics, thus leading to creating various techniques with titles such as *caution index*, *extended caution index*, *norm conformity index*, *individual consistency index*, and *optimal appropriate measurement*. Some of these indices directly use the test scores themselves, while others derive for Item Response Theory (IRT) or test for conformity to IRT model assumptions. Most of these statistics determine whether a given student's or group's scores conform to the typical pattern, but at least one (Tatsuoka and Tatsuoka, 1982) examines the consistency of a single person across multiple tests.

Why might person-fit statistics be needed?

For a variety of reasons, individual or group test scores may not accurately estimate ability. In presenting the argument for using person-fit statistics, Wright (1977) identified several types of behavior that cause invalid item responses. He noted types of students whose response patterns look askew. For example, he suggested that invalid item responses were given by

- ▶ people who got bored with a test and do poorly on later items,
- ▶ people who did poorly in the beginning because the test format confused them,
- ▶ people who never got to the later items,
- ▶ people who took wild stabs at the answers, and
- ▶ people who cheated.

Adding to this list in 1979, Levine and Rubin suggested that, on occasion, a particular student may be so unlike other examinees that the resulting test score cannot be regarded as an appropriate measure of ability. For instance, improperly aligning the answer sheet, using a poor test-taking strategy, or showing exceptional creativity in interpreting the question are

reasons for separating some students' scores from their counterparts'. In 1982, Levine and Drasgow also added these types to the list of mis-fit candidates:

- ▶ students who have high ability but who have had atypical schooling,
- ▶ lower ability students who copy answers from more able test-takers,
- ▶ those who are shown the answers to some items before taking the examination,
- ▶ students who are not fluent in English, and
- ▶ those who are conservative in using partial information.

They also argued that students who omit many test items are, in effect, taking a different test than those who answer all or almost all of the questions.

In more recent writings Levine and Drasgow (1987) contended that a "no response" to a test item should be treated as an option along with the usual choices, rather than as a response that is "not right." In later work (Levine and Drasgow, 1988), they also observed that students who deliberately fail a test often overused an ADADAD pattern while students who truly failed overused a BCBCBC pattern. While personality traits or response styles cause these aberrant patterns, Harnisch and Linn (1981) observed that the same number right on a test could mean very different things without reference to individual characteristics. For instance, on a 20-item test, a score of 10 can be obtained in 184,756 different ways.

They also contended that finding aberrant response patterns is no mere academic concern of psychometricians, rather, identifying individuals or groups with these patterns can reveal groups who have unusual instructional histories or individuals for whom the standard interpretation of the score is inappropriate. Removing the scores of these poorly assessed individuals would improve the accuracy of the aggregated group scores.

In researching this theory, Harnisch and Linn analyzed a set of test data from a state testing program that showed that different schools in different parts of the state had very different caution indices. They suggested that this variance could be caused by different curricula that didn't match the test, as well as by curricula in other parts of the state, but

they provided no empirical evidence to warrant choosing this alternative from among the several others that could also produce high-caution indices.

Earlier, in 1980, Frary argued that unusual test response patterns may be useful in identifying at least some types of test bias. But in 1982, in an effort to provide empirical evidence, Tatsuoka and Tatsuoka suggested that different patterns of aberrant responses were actually related to differences in instruction. In an early experiment, students were taught addition in signed-number operations by two different methods and were then given a test that contained both addition and subtraction problems. While the mean differences between the two groups were not significantly different, the person-fit indices were. In a second experiment, students who received a set of lessons based on using a different conceptual framework showed more aberrance than a group with consistent lesson frameworks.

In short, although the need for person-fit statistics has been documented and uses for it have been suggested, for the most part, it has not yet been applied to many settings. To date, the area has been largely one of potential, not of actual use.

Selecting appropriate fit statistics

Although many person-fit statistics already exist, selecting the appropriate ones has long been a problem. To date, much of the research has examined constancy across indices, rather than really investigating the ability to detect misfits. The need still remains to find "appropriate" person-fit statistics — those statistics that correctly identify mis-fit students, while, at the same time, making few false identifications.

Indeed, an early investigator complained that *"the trouble is that the formulas multiply not just like rabbits, or even guppies, but rather like amoebae: both by fusion and conjugation, and there seemed to be no general principle in selecting among them"* (Cliff, 1979). In

response, a number of studies have addressed the problem of selecting among statistics, typically using one or both of these two techniques:

- ▶ modifying real data to introduce aberrant patterns and having analysts look for differences among statistics, individuals, schools, or other categories of data; or
- ▶ generating simulated data with known characteristics and applying one or more sets of statistics to determine how well the misfitting subjects can be detected.

In 1981, Harnisch and Linn compared ten such statistics using data from the Illinois Inventory of Educational Progress, tests of reading and mathematics at the fourth grade. They found all but Kane and Brennan's Agreement Index to be strongly correlated.

Noting that the Harnisch and Linn study did not determine how well the techniques actually worked in finding mis-fitting students, in 1983, Rudner extended their research by making such a determination. He also included person-fit statistics derived from one- and three-parameter models of test design. Using data from the Scholastic Aptitude Test-Verbal and a General Biology Test, Rudner found that most statistics were strongly correlated, though the correlations were generally not as high as those found by Harnisch and Linn.

Commenting on the power of the statistics to find mis-fitting subjects, Rudner said that *"the proportions of correct identifications were not overwhelming."* Most of the statistics failed to identify more than 50% of the misfits, a notable exception being a weighted model fit statistic from a three-parameter model that identified about 75% of the misfits when 25% of the answers had been changed.

Frery (1982) reported less impressive results, although Frery's methods are not directly comparable to those of Rudner. Frery first compared three fit statistics,

- ▶ one from a Rasch model,
- ▶ Harnisch and Linn's (1981) modification of Sato's caution index, and

- one of his own devising using differences between the estimated probabilities of the most likely choices for each item and the choices actually made.

Frary's data came from twelve different tests of the same topic that were administered to different classes of students who took the same course from different instructors. All three indices were highly correlated, but none seemed particularly accurate. For the first two, expectations of mean scores could be calculated and, for both, the actual means were below the estimates. The scores on the modified caution index did not vary with the percent correct on the tests "to any meaningful extent." All three indices had weak correlations with test scores and aberrance seemed to be produced by aberrant responses to only a few items.

Since it had been suggested that the amount of preparation, personal problems, and the perception of fairness of the test could cause person mis-fit, questions to address these issues were appended to each test. However, no strong relationships emerged between the responses to these questions and the person fit-statistics.

Is fit a transitory or stable characteristic?

By applying fit statistics to successive tests, Frary (1982) also addressed the issue of whether fit is a transitory or stable characteristic. The correlations over time ranged from $-.21$ to $+.36$; thus, Frary concluded that "*These outcomes certainly do not speak for any substantial persistence of poor person-fit across successive testings for the tests and population studied.*" He speculated that his population — a largely white sample of college students — might have had a restricted range of ability at the high end of the ability scale. Frary concluded that the potential utility of person-fit statistics had not been demonstrated and urged caution in their use. However, the fit statistics Frary used are not among the more common indices.

Drasgow et al. (1987) also noted that correlating different indices of fit had limited use in determining which ones were best. Using simulated data, Drasgow et al. compared ten different indices in terms of their ability to detect aberrant responses under different degrees of aberrance and different ability levels. The indices varied considerably depending on the percent of mis-fitting data and the ability level of the simulated test-takers. Drasgow's Z3, the Rasch-derived F2, and one of the various caution indices that Tatsuoka developed, T4, proved to be the best, but they were not good enough: "...Z3, F2, and T4 provide rates of detection of some forms of aberrance that are nearly optimal but have inadequate rates of detection for other forms of aberrance." The three had nearly optimal rates when low-ability test-takers got high scores and when high-ability test-takers got low scores, but they did not perform well for average test-takers who had either spurious high or low scores.

In 1979, Drasgow and Rubin developed several person-fit statistics and applied them to simulated data. In 1980, Levine and Drasgow extended the research to show that the results for real and simulated data were comparable and that the statistics were robust with regard to errors in the estimation of ability. The Drasgow et al. (1987) study represented the culmination of a continuing interest in one line of research by Drasgow, Levine, and various collaborators. Among researchers of person-fit statistics, Drasgow et al. alone examined *appropriateness measurement* — their term for attempting to detect aberrant answer patterns — using polychotomous scoring.

In 1979, when they restricted their samples to respondents with low omission rates, Levine and Rubin observed much higher detection rates. Later, Drasgow et al. (1984) tried to remove this restriction by developing a model that considered non-responses as an additional, separate answer. They found that this polychotomous method was superior for detecting inappropriate patterns for low-ability respondents, but it was not as good as the dichotomous model for detecting high-ability respondents. Standardizing the statistics removed much of the ability-detection rate interaction.

That same year, Drasgow et al. (1984) conjectured that "*Excessively conservative examinees who are reluctant to use partial information, examinees who persevere on*

difficult items and other able, low scoring examinees with high nonresponse rates indeed do have inappropriately low number right scores." However, although Drasgow et al. used real data (they used SAT-Verbal items), they created the spurious response patterns simply by altering answers. That is, the "excessively conservative examinees" and other categories were hypothesized to exist, but no empirical evidence for their existence was presented.

While most person-fit research either presumed unidimensionality or tested for it, Drasgow et al. recently extended appropriateness measurement to multidimensional testing. One problem in appropriateness measurement is that none of the indices work particularly well on short tests. Since Drasgow et al. felt that they developed optimal indices for detecting aberrant patterns, their solution was to combine the data from several short tests of "distinct but correlated traits" into single tests. The measure that they felt was the optimal index, referred to as LRp, proved to be the best of nine such indices. However, it showed substantial variability between two unidimensional tests and among ability estimates. In general, the indices had better detection rates for simulated data from the Armed Services Vocational Aptitude Battery derived from *A Profile of American Youth*, than they did for real data taken from this study. Combining the data from a verbal test and a mathematics one increased detection rates.

Drasgow et al. noted several assumptions that may restrict the generality of the use of aberrance indices in real life. Their results assume that

- ▶ all items are equally likely to elicit spuriously high or low responses,
- ▶ all response patterns in the aberrant sample have the same number of aberrant responses, and
- ▶ the aberrant group has the same ability distribution as the normal group.

A later work by Drasgow and Levine (1986) extended the idea of optimal detection by determining maximum detection rates that they referred to as optimal detection. Their indices were "fairly high for spuriously high scores, but not for spuriously low ones."

Again, they found that scoring the responses polychotomously led to higher rates than dichotomous scoring.

In 1988, Levine and Drasgow noted that their previous work contained a logical flaw: If a technique detected aberrance, then, naturally, that form of aberrance was detectable. However, if a form of aberrance was not detected, one couldn't conclude that it was not detectable — some other procedure might work. Levine and Drasgow attempted to specify the power of the most powerful technique. To judge by the examples applied, this technique results in considerably better detection rates and is easier to compute.

In 1991, Reise and Due found that the Drasgow and Levine statistic had less power to detect aberrancy for short tests (tests with fewer than 20 items) than it did for long ones — an outcome consistent with the Drasgow and Levine research. However, it also had decreasing power as ability levels fell. This posed a problem since aberrancy is more likely to occur at low-ability levels. Reise and Due also concluded that the Drasgow and Levine statistic (and others) would not work well for peaked tests (tests that maximize reliability by restricting the range of item difficulties), nor would it work in the situation of adaptive testing. With adaptive testing, respondents are likely to be measured close to their ability range, the range where aberrance is most difficult to detect. Reise and Due, however, scored their data dichotomously and might have gotten superior results if they had followed the Drasgow and Levine recommendation to score polychotomously.

A common starting point for recent research in person-fit statistics is Sato's (1975) caution index. Tatsuoka and Linn (1983) expanded on it by applying Item Response Theory. They generated indices for one- and multiple-parameter models and used them to analyze person fit in a data set of elementary math test results (the grade was not specified). In addition, they applied Tatsuoka and Tatsuoka's Individual Consistency Index (ICI) to the data set. Both the ICI and one of the indices consistently identified children who used particular erroneous rules in answering math questions. Whether these indices could be generalized to other subject areas and other ages was not discussed.

In 1982, Tatsuoaka and Tatsuoaka developed both a norm-conformity index (NCI) and an ICI. On one hand, the NCI was derived from Cliff's (1979) research and detected deviations from the typical pattern of the group. On the other hand, the ICI was designed to measure consistency within an individual across several parallel tests. At one point, Tatsuoaka and Tatsuoaka suggested that the NCI might be used to detect differences in instruction, but the research they used for their example (Tatsuoaka and Birenbaum, 1979) was not conducted to test this hypothesis. In a related vein, Tatsuoaka (1984) used Item Response Theory to develop several "extended" caution indices. She hoped that such indices could be used to distinguish popular misconceptions about rules in mathematics from unusual ones, but she noted that *"caution must be taken when applying the indices to practical situations to spot atypical response patterns."* Again, applying these indices was limited to a restricted range of mathematics operations where explicit rules govern responses more than one might find in other curricular areas.

Unlike most person-fit indices where a high consistency score is unequivocally good, high scores on the ICI may be good or bad. A high ICI and a high test score indicate consistent right responses while a high ICI and a low test score suggest that a student is "hooked" on erroneous rules. The practical use of the ICI seems quite limited because it requires each student to take at least two, and preferably three or four, parallel tests. One of the caution indices that Tatsuoaka and Linn (1983) developed apparently behaved similarly to the ICI, but without the multi-test requirement.

While inter-index correlations were not the major focus of their research, Frary (1980) and Schmitt and Crocker (1984) also found strong intercorrelations among person-fit statistics using achievement test batteries on eighth grade students. Frary's methods — mostly uncommon ones — correlated less strongly and showed more variation from subtest to subtest than Schmitt and Crocker found using more common indices.

Finally, in this line of research, Yoes and Ho (1991) examined the percentage of students who were misfits on the reading, spelling, and science tests of the Stanford Achievement Test. The researchers chose these three subtests because the tests spanned the entire grade

range from third to twelfth grades. Using data from the Spring 1988 national norming sample, they used three person-fit statistics to determine what percentage of students fit the one-parameter Rasch model. In general, the percentages across grades and tests were low with a likelihood index giving somewhat higher rates of mis-fit (as much as 10%) than unweighted or weighted mean square fit indices. It would have been interesting to see this data disaggregated on a variety of demographic variables.

For somewhat different purposes, certain person-fit statistics have been developed to determine if a particular set of data fit a test design model, notably IRT models. Hattie and Rodgers (1987) used several statistics suggested by Wright and Stone (1979) and Wright et al. (1979) for determining fit: mean-square residual and total-t (analyses of item fit using these statistics as well as between-t was also a part of their research, but is not discussed here). In part, their research was motivated by concerns that Gustafsson (1980) expressed — removing people and items that do not fit a model produces only a spurious fit. (However, research by Chang and Bashaw (1984) on the 1977 norming sample of the Otis-Lennon School Ability Test found that removing mis-fitting students was detrimental to calibration results.) Rogers and Hattie's results for total-t are perplexing in that their data contained no guessing; more people were rejected when the data were two dimensional than when they were one dimensional. But when guessing was present, an increase in the number of people rejected in one-dimensional data resulted in similar numbers. In addition, as the data more and more violated the model, no commensurate increase in the number of rejections occurred.

The mean-square was insensitive to guessing, to heterogeneity in discrimination parameters, and to multidimensionality. It, too, failed to show the expected increase as the violation of the model became more extreme. Therefore, Rogers and Hattie concluded that using person-fit statistics could lead to accepting the model when it is "grossly inappropriate." They felt that excluding both people and items identified as mis-fitting provided no assurance of fit for the remaining items and people to the model.

In 1990, Reise reached a similar, though less strongly stated conclusion. Reise compared Bock's (1985) X2B, Drasgow's Z3, using simulated data and arranging a violation of the model in that the items were less discriminating for the data with violations than for the calibration sample. Reise found that the two statistics had many similarities, although X2B appeared biased towards rejection and reported that both statistics had an easier time identifying mis-fitting items than they did identifying mis-fitting people. In many cases, people were identified as fitting the model (that is, having an appropriate response pattern) when fit was computed using the incorrect item discrimination parameter.

In 1987, Kogut compared Drasgow's Z with a statistic that Molenaar (1987) developed. Kogut found the latter statistic to be more powerful, although it had problems. Kogut attributed the advantage of the Molenaar statistic to the fact that it is conditioned on the total score of the test rather than on a fixed ability and that it uses the exact distribution of the index rather than a normal approximation. The power of the index varied depending on whether the guessing occurred on easy or difficult items and depending on the ability of the guessers. These results would seem to point to many unresolved issues concerning the various kinds of purported test-takers. Although Wright (1977) named a variety of traits or responses styles that would differentiate different individuals, no empirical investigations have been made into the actual existence of these styles. Except for Kogut's variation of which items were guessed at, no investigations have determined if different person-fit statistics are differentially effective for these different types of aberrance.

Similarly, in 1990, Molenaar and Hoijtink argued against fit statistics based on fixed ability. They also observed that a response pattern that is aberrant for a person of low ability may not be aberrant for a person of high ability. They proposed that ability intervals be created in a fashion similar to Drasgow's proposal in 1985. They concluded that determination of aberrance should be made only by comparing the value of a person-fit statistic of an individual to another of the same ability.

More recently, Klauer (1991) developed another fit statistic which he compared to Molenaar and Hoijtink's (1990). He observed that the Molenaar and Hoijtink statistic and

his own were differentially sensitive to different types of aberrant responses and that the Molenaar and Hoijsink statistic was not sensitive at all to one type of violation. He observed, though, that *"it may be premature to draw conclusions concerning the relative merits of both indices for detecting deviant responses patterns and for diagnosing factors causing such deviations in practical testing."*

How has person-fit statistics been applied?

While most research efforts in the person-fit area have been directed toward theoretical and methodological concerns over the nature, accuracy, and interchangeability of person-fit statistics, some researchers attempted to practically apply them as well. For instance, in 1980, Frary calculated four different fit statistics on a large sample of eighth graders who took a commercial achievement test battery. Along with the moderate to strong intercorrelations mentioned earlier, Frary found that on some tests blacks and females differed from whites and males, respectively. Overall, females showed fewer aberrant responses than males, but racial differences occurred in both directions. Among low-scoring students, the effects were consistent: whites and females had fewer unusual choices. These findings raised the possibility of test bias. However, using a "knowledge assurance" statistic proposed by Horst (1966), Frary concluded that blacks and males were better at correctly guessing items for which they had only partial knowledge than were whites and females.

In another case in 1981, Doss applied a residual mean square statistic from the computer program RASCH in the PRIME system to a fifth grade Chapter 1 setting where children were given the Iowa Tests of Basic Skills. He examined the effect of removing the poorest fitting 10, 20, and 30 percent of students on the accuracy of the pre-test predictions. Although the N dropped substantially as more and more students were removed, accuracy of prediction increased with each removal. While such an improvement in accuracy is interesting, there is some question that Doss's setting provides a meaningful measurement situation. The test is badly matched to students' abilities. Even though some students (Doss did not specify how

many) took the fourth grade level of the battery, 25% of the students scored at or below the chance level. After the worst fitting 30% of students were removed, only 13% of the Chapter 1 students remained. Finally, the students showed losses on the tests from pre-test to post-test. The person-fit measure could be used to document that the testing situation is not meaningful, but it seems like an elaborate procedure for that purpose, although its objectivity might serve the purpose well.

In still another study in 1984, using a variety of indices and the Metropolitan Achievement Tests in reading, mathematics, and science in seventh and eighth grades, Schmitt and Crocker investigated the relationship between scores on the Test Anxiety Scale for Adolescents and person fit. Students in the middle-ability range showed no relationship between test anxiety and person-fit indices. High-ability, low-anxious students showed greater mis-fit than high-ability, high-anxious students. At the low-ability end, the reverse was true: low-ability, low-anxious students showed less mis-fit. The authors offered some conjectures on the findings in terms of a Cognitive-Attentional Theory of Test Anxiety, but presented no data to support their notions.

In an undated paper, Westfall and D'Costa reported using a Rasch fit statistic and a statistic from IPARM (which they stated is similar to Sato's caution index) to examine improvements in classifying and predicting success for Air Force Academy freshmen. All freshmen took a French test and were classified as being low, intermediate, or advanced. Predictions were calculated for freshmen grades using raw scores, Rasch, and IPARM indices. The Rasch statistic improved prediction and classification slightly for low and intermediate students, while the IPARM index worked somewhat better. Neither index increased the prediction or classification accuracy for advanced students.

In another example, Doss and Ligon (1985) applied person-fit statistics to solve a specific problem in which some students had been inadvertently given the wrong form of a test. Their decision rules seem reasonable, but as Doss and Ligon noted, in this particular situation, they had no empirical way of knowing their error rate. In any case, this study was

an application that does not seem to further the general utility of the technology and one hopes that the need for it does not arise often.

Observations

Nearly twenty years after Sato introduced his caution index, person-fit statistics still seem to be in the realm of the potential. After reviewing the literature, three strong impressions can be drawn:

- ▶ The research has been largely unsystematic,
- ▶ The research has been largely atheoretical, and
- ▶ The research has not explored whether the potential of person-fit statistics can actually be realized in applied settings.

We consider each impression in turn.

The research has been unsystematic.

To date, the research has involved school-aged children, college students, and simulated respondents. It has looked at differences by ethnicity, sex, and degree of test anxiety. But, taken together, these examinations do not constitute a substantial body of knowledge. It is suggestive, not definitive.

For instance, Wright (1977) theorized about a variety of test-taking traits. If these traits or styles actually exist, they should produce different patterns of aberrance. To date, however, all aberrance detection has simply been in terms of spuriously high or spuriously low scores or in terms of characteristics of the test, such as length or range of item difficulty. We need more refined, systematic studies to determine if different statistics are required to detect different traits.

Similarly, some recent research strongly suggests that different statistics will be useful for detecting different kinds of aberrance, but this suggestion must be followed up systematically. Also, recent studies have contended that the detection of aberrance should be restricted to students of the same ability group — a seemingly reasonable suggestion, but much of the research did not limit itself in this way nor has anyone specified how similar "same ability" students should be. A systematic retrospective look is necessary to determine how ability considerations might have affected this research and more systematic studies must be undertaken to determine if ability considerations really are important in comparing patterns of responses. Many studies have set the aberrance rates at 15%, 25%, or 30%, finding, as one would expect, higher detection rates for higher rates of aberrance. But the one study that examined aberrance rates in an existing achievement test found no more than ten percent mis-fits with any of the various indices used; usually considerably less were found. Although aberrance in this instance was deviation from a specified model, this finding raises questions about the utility of the various indices. Although there has been discussion of "optimal" rates of detection, no one has yet considered whether optimal is good enough in a practical setting.

The research has been atheoretical.

The research has not incorporated recent advances in cognitive psychology nor has it dealt with any theoretical considerations. The various studies mention *ability*, but it is an undefined construct in some research and a construct that is narrowly defined by IRT technique in others. The ability parameter has been estimated from the SAT, the GRE, and various achievement test scores, but it has not been defined as a meaningful construct outside of such parameter estimation.

The research has not explored practical applications.

Much of the literature mentions the potential of the indices described and says that they may be useful in one way or another, but the necessary follow-up studies must be undertaken to determine if the potential can be realized. We need to address these questions:

- ▶ Do differences among groups reflect differences in curriculum or attendance?
- ▶ Do individuals who omit many items need counseling in how to take a test?
- ▶ Do spuriously high scores reflect cheating?
- ▶ Do different ethnic groups consistently test differently and, if so, what does that mean? Does ethnicity or social class produce the differences (Some research has found class to be the more potent variable (Hodgkinson, 1991).)?

In general, we need more clinical, practically oriented studies that find aberrant patterns of responses and then follow up with the respondents. No one has empirically investigated what these respondents are like. Can anything meaningful be said about them beyond the fact that they do not look like typical respondents?

In short, although the need for person-fit statistics has been documented and uses for it have been suggested, for the most part, it has not yet been applied to many settings. To date, the area has been largely one of potential, not of actual use.

What fit statistics have been proposed?

Rasch Model Approaches

Wright and Panchapakesan (1969), Mead (1976), and Mead and Kreines (1980) discussed two statistics that can be used to evaluate whether the observed number of correct responses

agrees with the number predicted from the Rasch model parameter estimates. The first is an unweighted total fit mean square:

$$U1_i = \sum_{j=1}^N \frac{(u_{ij} - P_{ij})^2}{P_{ij}(1 - P_{ij})} / N$$

where i indexes the examinee, j indexes the N items, P_{ij} is the probability of a correct response predicted by the Rasch model, u_{ij} is the observed item response.

The second is a weighted total fit mean square:

$$W1_i = \sum_{j=1}^N (u_{ij} - P_{ij})^2 / \sum_{j=1}^N P_{ij}(1 - P_{ij})$$

The two person fit statistics, which are computed for each individual by summing over items, are differentially sensitive to different kinds of items. The unweighted fit statistic is influenced more by very hard and very easy items, while the weighted fit statistic is more sensitive to items of near mean difficulty. The weighted fit statistic is incorporated in the BICAL computer program (Wright et al., 1979) as part of the standard printout.

Birnbaum Model Approaches

Rudner (1983a) first suggested replacing the probability of a correct response predicted by the one parameter Rasch model with the probability of a correct response predicted by the three item parameter Birnbaum model - the model used by NAEP. Two statistics, denoted $U3_i$ and $W3_i$, are obtained by substituting the probability of a correct response given by the Birnbaum model for the P_{ij} values in the two previous equations.

A third Birnbaum model approach is based on the likelihood function, $L(\theta_i)$, which is the product of the probabilities of the observed item responses:

$$L(\theta_i) = \prod_{j=1}^N P_{ij}^{u_j} (1 - P_{ij})^{(1-u_j)}$$

This function indicates the compound probability of the observed response pattern for a given ability estimate. Following the suggestion of Levine and Rubin (1979), the log of the likelihood function is evaluated at the value of θ_i which maximizes the likelihood equation, that is, at the maximum likelihood estimate of ability. The log of the likelihood function can be rescaled to a 0-1 metric by taking the geometric mean of the probabilities.

Correlation Approaches

Two correlations of an individual's 0-or-1 item response pattern with the corresponding group determined p-values can be used to indicate the tendency of an examinee to correctly answer easy items and miss the hard ones. The first correlation is simply the personal point-biserial correlation, r_i . The second is the personal biserial correlation, rb_i , under the assumption of a normally distributed continuous variable underlying u_{ij} (Donlon and Fischer, 1968).

Sequence Approaches

Tatsuoka and Tatsuoka (1982) and Harnisch and Linn (1981) discuss two approaches based on the responses to items sorted from easiest to hardest based on p-values. The "ideal" response pattern would consist of a string of correct responses (1's) followed by a string of incorrect responses (0's). High ability examinees would have a long string of 1's. Low ability examinees would have a short string of 1's. Examinees with odd response patterns would tend to miss some easier items and correctly answer some more difficult ones. Their response patterns would not form a consistent series.

Tatsuoka and Tatsuoka's (1982) norm conformity index (NCI_i) is computed by taking the dominance matrix from the ordered item response vector. The elements of the matrix, n_{ij} , equal 1 when the examinee gets item i wrong and item j right. Otherwise $n_{ij} = 0$. The NCI_i is then:

$$NCI_i = 2 S_a / S - 1$$

where S_a is the sum of the above-diagonal elements, S is the sum of all matrix elements. Harnisch and Linn (1981) have stated that the NCI_i is mathematically related to van der Flier's (1977) U' statistic.

A modified caution index, C_i , based on Sato's (1975) caution index, was proposed by Harnisch and Linn (1981). This statistic is computed as:

$$C_i = \frac{\sum_{j=1}^{n_i} (1 - u_{ij}) n_j - \sum_{j=n_i+1}^N u_{ij} n_j}{\sum_{j=1}^{n_i} n_j - \sum_{j=N+1-n_i}^N n_j}$$

where n_i is the total score for examinee i , n_j is the number of correct responses to item j .

Discussion

There have been numerous discussions and evaluations of fit statistics in the literature (Andersen, 1973; Drasgow and Levine, 1986; Gustafsson, 1980; Haertel, 1989; Harnisch and Linn, 1981; Levine and Drasgow, 1982; Levine and Rubin, 1979; McKinley and Mills, 1985; Reise, 1990; Rudner, 1983a; Smith, 1991; Wainer, Morgan, and Gustafsson, 1980). Many of these studies used large simulated datasets in order to control the characteristics of the dataset and test for power. The observations have been quite consistent despite the differences in data generation and analysis.

Possible criteria for the selection of a technique include power in the intended application, availability of item parameters, computation requirements, and cost. The most generally applicable statistics are the norm conformity index and the modified caution index. These statistics are not, however, the most efficient statistics. With shorter classroom tests, the point biserial correlation and the personal biserial correlation are among the most efficient approaches. These statistics have the advantages of requiring only modest numbers of items and examinees, being simple to compute, and being easy to communicate. For larger assessments such as NAEP, statistics based on the Birnbaum model and the Rasch model unweighted fit statistic tend to identify the most examinees with spurious total scores.

With very few exceptions, the effectiveness of the approaches increases as the number of examinees with spurious scores. The likelihood statistic, $L(\theta_i)$, works well in identifying examinees with spuriously high and low total scores. The weighted Rasch model fit statistic, $W1_i$, follows a similar trend, but does not do as well with spuriously high scores.

The unweighted Rasch model fit statistic, $U1_i$, and the two correlational approaches, r_i and rb_i , tend to work well on shorter tests. The opposite trend occurs with the unweighted Birnbaum model fit statistic, $U3_i$. This approach works well with longer commercial tests, but poorly with shorter classroom tests.

The two sequence approaches, the norm conformity index (NCI_i) and the modified caution index (C_i), tend to identify comparable proportions in all situations. These two approaches are among the most stable of the statistics.

We view the weighted fit mean square based on Birnbaum's 3 parameter logistic model as the approach most applicable to NAEP data and used it in this study. It

- incorporates the same measurement model as NAEP,
- is most influenced by items of median difficulty,
- has a standardized distribution, and

- has been shown to be near optimal in identifying spurious scores at the ability distribution tails.

The technique is not without its limitations. Wainer, Morgan and Gustafsson (1980) argued that fit statistics based on unconditional maximum likelihood do not have known asymptotic properties. Smith (1991) noted that the observed and expected response for an item are not entirely independent since the observed response was used in computing the item parameters and the ability estimate. In addition, the asymptotic theory is based on two continuous variables and the observed response is not a continuous variable. The effect of these limitations is that the variance of the fit mean square may not be accurate. Because of the intent of this research, the size of the NAEP dataset used in this study and our use of fit as a relative rather than absolute indicator, we feel these limitations are minor and will have a limited effect, if any, on our analysis.

It should be noted that the weighted fit mean square has an expected value of 1.0. Fit greater than 1.0 indicates poorer than expected fit of the respondent to the model. Fit less than one indicates fit that was better than expected. For example, suppose we had a test of 10 items and the probability of a correct response to each of the items was .7. We would expect a respondent to get a total of 7 items right and this would correspond to a weighted fit mean square of 1.0. If we examine each item individually, we would expect each item to be correct because the probability of a correct response is closer to one than zero. Thus, with this hypothetical example, if a person obtained a total of 8 items right, fit is better than expected and the fit statistic is less than 1.0 (0.81). If the respondent got 6 items right, fit is worse than expected and the fit statistic is greater than 1.0 (1.19).

What are the implications for NAEP?

Traditional psychometrics are based on a large number of individuals responding to a common set of items. Individuals are treated as a sample from some larger population, but the test is not. Content is treated as a fixed effect. Except for occasional non-responses and

other minor problems, such a formulation provides a complete data set. Total scores can be computed for individuals and correlated with background and attitude questions. With such a model, trends and relationships in the data can easily be explored.

The intent of large scale assessments, such as NAEP, is to generalize beyond the particular exercises administered. Both content and persons may be conceived as random effects. When the intent is to learn about a population's proficiency with respect to a domain of exercises, the efficiency of the data collection can be improved by sampling more items from the domain, and giving each of those items to fewer individuals. With this approach one ends up knowing relatively little about individuals, but a great deal about the mean percent correct on a potentially larger set of items. Average performance could be computed for the nation as a whole as well as for sub-populations of interest.

Such matrix sampling has characterized NAEP since its inception 20 years ago. As originally employed, the model had the added advantage of limiting the potential for invidious comparisons -- a major concern at the time. The problem with the original approach, however, soon became apparent. If you wanted to know how the nation was doing in mathematics, you ended up with 450 different answers - one for each item.

NAEP's original solution to this problem was to organize NAEP exercises into a limited number of test books. By randomly assigning books to individuals, mean percent-corrects could be reported over these sets of items. This, however, limited trend analysis to specific and unchangeable sets of items and provided little information about skill distributions.

Building on refinements in item response theory, ETS introduced scaling procedures into the national assessment. Scale scores are computed using IRT models modified to handle multiple matrix sampling. In this way, scale scores in the domains of interest can be reported without depending on specific items. The methodology depends only on the development and maintenance of a bank of calibrated exercises in each content domain. New exercises can be added to the assessment as old ones are released to the public without compromising in any way the ability to measure change. Furthermore, results of the

assessments can be reported directly in terms of the skill areas of interest rather than in terms of specific exercises, whose coverage of the skill area is incomplete and whose psychometric characteristics are unknown.

The conventional machinery for latent trait estimation was well equipped to handle the complexities created by BIB spiraling (NAEP's variant of multiple matrix sampling). Since exercise-examinee sampling procedures dictate that any one student only takes a very small number of exercises per skill area, the estimation of individual latent trait levels is very imprecise. Since the assessment is mainly interested in population values, this imprecision could generally be tolerated. In order to estimate the measurement error in aggregate estimates, especially when large numbers of presented items are omitted, ETS introduced the concept of "plausible values," based on Rubin's work in data imputation. The Bayesian method used to calculate the plausible values also reduces the variance of the individual level estimates.

It is important to note several cautions that must accompany adoption of an IRT approach (Lord and Novick, 1968). In theory IRT item parameters are invariant (Lord and Novick, 1968; Rudner, 1983b). In practice, however, the extent of agreement between scaling of alternative uses of the same item is frequently much lower than expected. This may be due to context effects, to failure of unidimensionality assumptions, to instabilities in estimation, failure of local independence assumptions or to other sources. A second concern is that items that are otherwise fully acceptable are occasionally dropped from use because they do not "fit" an IRT model. A final concern is that the estimation of the item and examinee parameters under more sophisticated IRT models is sometimes unstable due to "local minima" problems.

Within the NAEP framework there are numerous technical issues which can be illuminated by assessment accuracy statistics. We describe some of the many issues and indicate how the use of assessment accuracy statistics can address those issues.

Describing performance - Group results for NAEP are currently described in terms of behaviorially anchored averages and standard errors. The primary interest has been in trends and group differences. Implicit in this reporting is that all group statistics and all assessments are extremely accurate. Yet, there are indications that this assumption may not hold. Different items and item types, differences in culture, and differences in maturation, for example, can be expected to affect assessment accuracy. Fit statistics have the potential to identify when these differences affect NAEP results.

Evaluating innovative item types - NAEP has been a leader in the current move toward innovative item types. While some will argue that such items are more valid than traditional multiple choice items (Wiggins, 1990), the question has not been examined empirically. It may be that students try their best on multiple choice questions but refuse to exert the effort needed to respond to open-ended questions. Fit statistics can help evaluate whether the innovative items function as intended.

Item bias - Item bias can easily be evaluated for NAEP using any of a wide variety of bias detection techniques (see Rudner, Getson, and Knight, 1980 for a discussion of several approaches). In addition to the traditional use of race, gender, and region as grouping variables, one could use other relevant demographic characteristics, such as curricular differences. Most importantly, grouping can be made more homogeneous by trimming based on fit statistics. For example, a racial bias analysis would benefit by including only white examinees with small person-fit statistics (i.e. majority examinees with good fit) and Black examinees with large person fit statistics (i.e. poorly assessed minority examinees). Flagged items can be analyzed and organized to serve as positive and negative models for future item developers.

Plausible values - Parameter estimation in item response theory usually requires that a large number of items be administered to each examinee. As the number of items is reduced, errors of estimation increase. Because some examinees respond to very few cognitive items in NAEP, traditional estimation procedures can produce unacceptably biased estimates of examinee ability. Imputation methods outlined by Mislevy (1988) were introduced as a means

to correct this problem starting with the 1984 assessment. Ability estimates are iteratively imputed through a model incorporating background variables. The process raises several basic questions that can be addressed through an evaluation of assessment accuracy. What is the increase in precision introduced by imputation? Are heavily imputed values as consistent with the model as less heavily imputed ability estimates?

Motivation - Motivation can play a major role in the National Assessment. Students knowing that they have nothing to risk may not try to do their best, thereby compromising the assessment. The 1987 Technical Review Panel observed that one of the major threats posed to the National Assessment program by the State program is a possible change in the motivational context of the assessment. If the students in one state try to outshine the students in another, their activities may compromise the validity of the results. Motivational problems are reflected in fit statistics. The student that is not trying his or her best is likely to have a poor fit; the highly motivated student is likely to have high fit.

Design effects - One observation made in investigating the reading anomaly was that the scores on reading booklets preceded by computer science booklets were lower than those that were preceded by booklets in other subject areas. If computed on a routine basis, fit statistics could flag such problems early in the analysis of the assessment results.

Chapter 3: 1990 Trial State Assessment

Methods

This methods section describes the trial state assessment data used and how we computed person fit.

Data

In February and March 1990, NAEP conducted a Trial State Assessment of mathematics knowledge, skills, and understanding of representative samples of eighth grade students in public schools. Thirty seven states, two territories, and the District of Columbia voluntarily participated in the assessment. Each state had approximately 2,500 respondents.

In the Trial State Assessment, there were seven test booklets, numbered from eight to fourteen. Each booklet contained common and mathematics background items and three blocks of mathematics items. Items from all five subscales appeared in each block. The booklets were constructed in such a way that each mathematics item block appeared in three booklets and every pair of blocks appeared together in exactly one booklet. Booklets were then spiraled and administered. For further information on how the Assessment was administered, the reader is referred to *The NAEP 1990 Technical Report* (Johnson & Allen, 1992).

We analyzed a randomly selected subset of the data in the Secondary-Use Data Files. Each student's record consisted of responses to three blocks of mathematics items, common background items, mathematics background items, the students' teacher's responses to a questionnaire, and the students' school administrator's responses to a questionnaire. Each record also contained plausible value ability estimates for each subscale and a composite.

To obtain representative samples of eighth grade public school students, NAEP sampled approximately 100 schools in each of the 40 jurisdictions, stratified on the basis of urbanicity (central city, suburban, or other), median household income, and, in states with significant minority populations, minority enrollment. From each selected school, a random sample of 30 students was drawn for a target sample of about 3,000 students from each jurisdiction.

The public use datatape contains over 100,000 records, one for each respondent, and each record is 1,699 characters wide. For these analyses we took a more manageable random sample of 2,200 respondents. Sampling weights were adjusted to keep the sum of the weights for the respondents in each state the same as in the original dataset. Thus, the subsamples were state representative. This fact is confirmed in an analysis described later.

Computation of fit

In order to compute weighted total IRT model fit mean square statistics, we needed estimates of the probability of a correct response, P_{ij} , and the observed response u_{ij} , for each individual. P_{ij} was computed using the 3 parameter logistic model, estimates for subscale θ 's that appear in the public use dataset, and item parameters appearing in the technical documentation. Values for u_{ij} were taken from the public use dataset. Subscale thetas were used rather than the composite theta because the latter is simply the average of the subscale thetas and not a model derived estimate of overall mathematics ability. A total of 13 mean square fit statistics were computed - one for each of the 5 scales, one for each of the 7 cognitive item blocks, and one across all applicable cognitive items. As with the ETS analysis, omissions were coded as incorrect responses.

Results

This results section describes

- ▶ the overall fit of individuals,
- ▶ fit by state,

- ▶ fit by key demographic variables,
- ▶ fit by item block order,
- ▶ the relationship of person fit to proficiency,
- ▶ prediction of misfit, and
- ▶ the effect of trimming misfitting students

using the 1990 Trial State Assessment data. The appendices contain further analysis by subtests.

Description of Fit Statistics

The distribution of the IRT model-fit mean square statistic computed over all item responses is shown in Figure 1. The distribution follows the familiar bell curve quite well and is extremely peaked around the expected value of 1.0. The mean is .97 and the standard deviation is .17. Fit for the NAEP Trial State Assessment is quite good and there are very few respondents with extremely abnormal response patterns.

Figure 1
Distribution of Person Fit in NAEP TSA

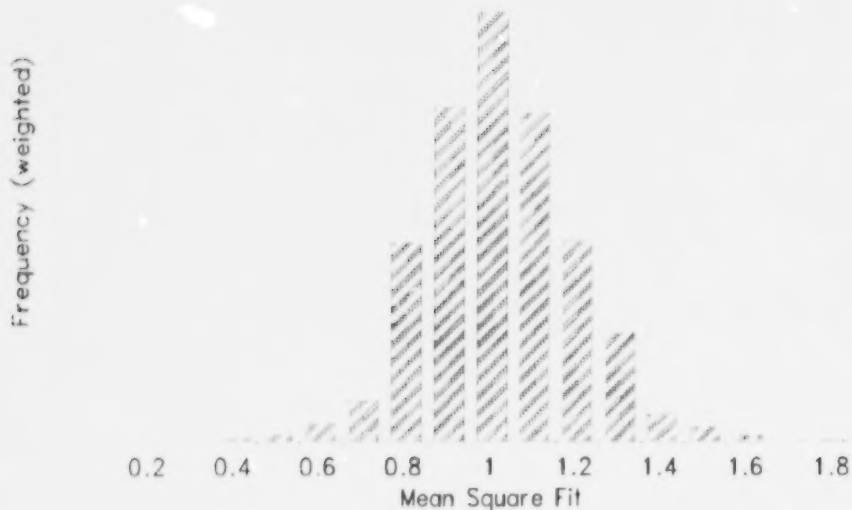


Table 1
Descriptive Statistics of Fit Mean Square Statistics

Subscale/Block	Mean	SD	Skew	50	Pctiles 70	90
Numbers & Oper	.96	.29	.56	.94	1.08	1.32
Measurement	.96	.41	.99	.89	1.12	1.49
Geometry	.91	.30	.43	.89	1.05	1.31
Data Anl, St, Prb	1.14	.58	.77	1.07	1.39	1.92
Algebra & Functns	.93	.41	.90	.88	1.09	1.45
M3	.92	.29	.23	.92	1.05	1.25
M4	.95	.26	.73	.93	1.06	1.28

Subscale/Block	Mean	SD	Skew	50	Pctiles 70	90
Numbers & Oper	.96	.29	.56	.94	1.08	1.32
M5	1.10	.31	.55	1.10	1.23	1.46
M6	.95	.38	3.17	.90	1.07	1.37
M7	.95	.28	.66	.92	1.08	1.30
M8	.99	.37	.82	.95	1.15	1.50
M9	.95	.26	1.60	.92	1.04	1.28
Overall Fit	.97	.17	.25	.96	1.05	1.19

The IRT model-fit mean-square statistic was also computed separately for each examinee on the NAEP Mathematics subscales, the mathematics item cognitive blocks (M3-M9), and over the entire assessment. Among the subscales and item blocks, the distributions of fit statistics, shown in Table 1, were highly consistent. The means tended to be just a bit less than the expected value of 1.0. The standard deviations ranged from approximately .2 to .6. Because the statistic is bounded by zero at the lower end but unbounded at the upper end, the distributions were all positively skewed.

The distribution of fit statistics of the Data Analysis and Statistics subscale and item block M5 showed a higher degree of misfit than any of the other subscales or blocks. Thirty percent of the students had fit statistic values higher than 1.39 and 1.23, respectively. We will offer an explanation for the higher degree of misfit on the Data Analysis and Statistics subscale and M5 block later in this paper.

Fit by State

Were there state differences in average mean square fit? An affirmative answer would imply an unintended bias in the TSA data - some states would have elevated or, more likely, suppressed scores due to a greater number of examinees with poor fit. By removing or "trimming" these poorly assessed individuals from the analysis, more accurate state estimates

could be obtained. As shown in Table 2, however, fit by state, like the overall fit, was exceptionally good. Fits varied from .88 to 1.06 with little within state variation. Latter we show that trimming 10% of the sample with the poorest fit has little effect on the results.

Table 2
Average mean square fit (and standard deviation)
on the NAEP TSA by state

AL	.95	(.16)	MN	.98	(.13)
AR	.98	(.18)	MT	.98	(.15)
AZ	.95	(.12)	NC	.92	(.16)
CA	1.05	(.13)	ND	.92	(.16)
CO	.98	(.18)	NE	1.00	(.24)
CT	.97	(.21)	NH	1.00	(.16)
DC	.93	(.15)	NJ	.96	(.16)
DE	.99	(.20)	NM	.93	(.16)
FL	.88	(.16)	NY	.98	(.16)
GA	.94	(.14)	OH	1.00	(.17)
GU	.98	(.16)	OK	.96	(.16)
HI	.94	(.20)	OR	1.01	(.17)
IA	.96	(.17)	PA	1.06	(.17)
ID	.96	(.17)	RI	.92	(.17)
IL	1.00	(.18)	TX	.98	(.17)
IN	1.00	(.18)	VA	1.00	(.21)
KY	.96	(.16)	VI	.95	(.20)
LA	1.00	(.15)	WI	.98	(.15)
MD	1.00	(.17)	WV	.97	(.14)
MI	.95	(.18)	WY	.99	(.16)

Fit by key demographic variables

NAEP regularly reports proficiency levels for a number of population subgroups, designated according to gender, race, type of community, region, parents' education, and other variables. If these groups differ in average fit, then there may be a significant bias in traditional NAEP reporting. As shown on Table 3, there are no meaningful differences in means or standard deviations.

Table 3
Means (and Standard Deviations) of Fit Statistics
by Select Demographic Characteristics

Group	Overall MS Fit	Group	Overall MS Fit
Nation	.971 (.173)	Region	
		Northeast	.984 (.186)
Race/Ethnicity		Southeast	.955 (.166)
White	.972 (.178)	Central	.974 (.172)
Black	.967 (.155)	West	.974 (.165)
Hispanic	.976 (.148)	Community	
Asian	.970 (.178)	Ext rural	.953 (.162)
Amer Indian	.948 (.166)	Adv. Urban	.952 (.155)
Sex		Disad Urbn	.976 (.188)
Males	.980 (.179)	Other	.975 (.177)
Females	.961 (.165)		

Fit by Item Block Order

A question that has often been raised in designing tests and large scale assessments concerns testing time. The longer the testing time, the more items can be administered, and testing quality can increase. Up to a point. After some period of time, students can be expected to be fatigued and no longer try.

To test for a fatigue effect, we computed the mean fit by block order. Table 4 below shows mean fit statistics for each item block according to whether the block appeared second, third, or fourth in the booklet (the first block in each booklet always consisted of background questions). The numbers in parentheses indicate the item block number (M3-M9). It should be noted that these means are based on all of the items in a block, regardless of which subscale they represent. The far right column indicates the mean fit statistic value for the Data Analysis subscale for each booklet.

As shown in Table 4, there are no differences in the average fit of the first, second, or third block in a test booklet. There was no detectable block order or fatigue effect.

Table 4
Mean Fit Statistics of Booklets According to
Block Order and Data Analysis Subscale

Booklet	2nd Block	3rd Block	4th Block	Data Analysis
8	.90 (M3)	.94 (M4)	.95 (M6)	.88
9	.96 (M4)	1.09 (M5)	.98 (M7)	1.37
10	1.09 (M5)	1.01 (M6)	.99 (M8)	1.39
11	.88 (M6)	.93 (M7)	.97 (M9)	.94
12	.95 (M7)	1.02 (M8)	.93 (M3)	.98
13	.97 (M8)	.92 (M9)	.95 (M4)	.94
14	.96 (M9)	.95 (M3)	1.13 (M5)	1.54
Mean	.96	.98	.98	1.14

Block M5 and the Data Analysis subscale

Table 4 also provides insight into the higher means square fit for item block M5 and for Data Analysis noted above. The mean fit statistic value for item block M5 is higher than any other regardless of the order in which it appears in the booklet. As the same time, the mean fit statistic values for the Data Analysis subscale are higher for Booklets Nine, Ten and Fourteen. For the other four subscales, there were no differences between mean fit values across the seven booklets. One may infer then that the high mean values for item block M5 come from the Data Analysis subscale.

So what is unique about item block M5? Item block M5 is the only block composed entirely of open-ended items. That alternative item formats may have an impact on person misfit is certainly an interesting and important issue. Clearly this relationship warrants further analysis.

Relationship of Person Fit to Proficiency

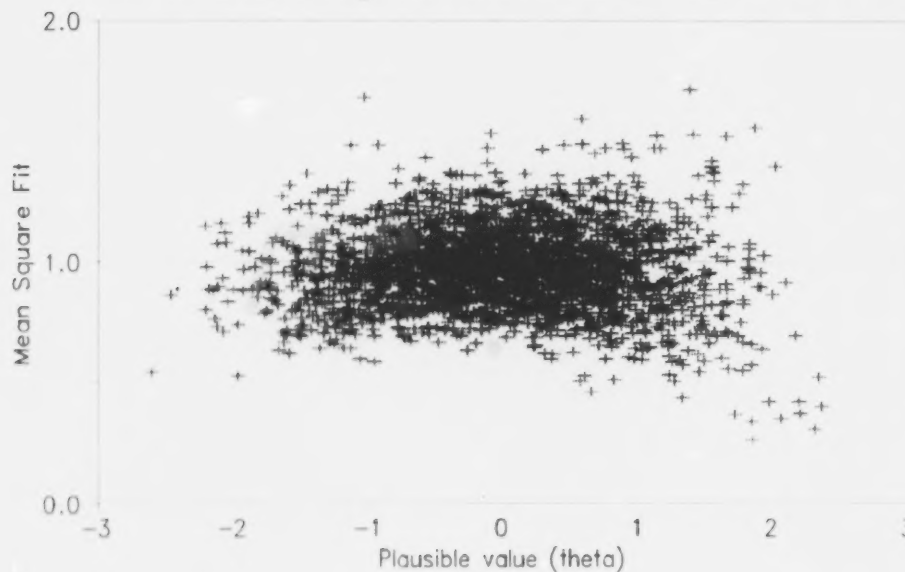
Is there a relationship between person fit and ability? One might expect large numbers of exceptionally capable and exceptionally poor students not to care about the assessment, not try, and consequently have poor fit.

Table 5 displays the correlations between model fit statistic values and plausible values for each of the subscales and composite. These correlations ranged between $-.08$ and $-.01$, indicating essentially no relationship between proficiency and model fit. The scatterplot of overall ability and fit, shown in Figure 2 confirms the lack of relationship. Scatterplots for the subscales show the same lack of a relationship.

Table 5
Correlations Between Fit Statistics and Plausible Values

Numbers & Operations	$-.02$
Measurement	$-.08$
Geometry	$-.07$
Data Analysis	$-.08$
Algebra & Functions	$-.01$
Overall	$-.02$

Figure 2
Relationship between theta and fit



One important question concerns the degree to which model fit is related among subscales and item blocks. That is, if a student shows model misfit on one subscale or item block, is there a tendency toward misfit on other parts of the assessment. Pearson correlations of model fit statistic values between subscales range between .01 and .11, suggesting no relationship. The correlations of model fit between item blocks ranges from -.20 to .17, also suggesting little or no relationship.

Correlation coefficients are misleading in that they examine relationships across an entire range of values, whereas we are interested in whether misfits on one scale also tend to be misfits on other scales. To examine whether students that show poor fit on one scale also show poor fit on other scales, we identified the percent of students whose mean square fit was 1.2 or greater on each scale were also misfits on the other scales. The results are shown on Table 6. For example, 26.2% of the students that had poor fit on the Numbers and Operations Scale also had poor fit on the Measurement Scale.

Table 6
Percent of poorly fitting students
on each scale that are also poor fit on the other scales

Selected for poor fit on	Percent of poor fitting students					Overall
	No&Ops	Measnt	Geomtry	Data Anal	Alg&Fns	
No & Ops	100.0	26.2	21.2	41.7	30.1	28.8
Measurement	19.2	100.0	18.7	41.3	24.4	20.7
Geometry	20.3	24.7	100.0	37.4	20.6	22.6
Data Analysis	17.6	23.8	16.4	100.0	21.2	15.3
Alg & Fns	24.0	26.5	17.7	39.9	100.0	25.4

From Table 6, it can be seen that there is typically a one in five chance that if a student has poor fit on one scale, he will also have poor fit on other scales, excluding Data Analysis. The odds double with regard to Data Analysis, the scale with the poorest overall fit and the scale with open-ended questions. From Table 6, it is apparent that few students with poor fit on one NAEP scale tended to have poor fit on the other scales. There is only a slight relationship of fit across NAEP scales.

Prediction of Person Misfit

In addition to the cognitive mathematics items, each assessed student was administered a demographics questionnaire and a mathematics background questionnaire. NAEP also gathers information from the students' teachers and information concerning school characteristics and policies. We sought to identify relationships that might exist between misfit and these demographic characteristics.

Pearson correlations were computed between subscale fit statistics and a variety of variables. Only a handful of variables had correlations larger than .10 or less than -.10. The majority of these correlations were near zero. None had an absolute value greater than .15.

A series of discriminant analyses were conducted to identify the variables that best predict which examinees have poor fit. We categorized the students into two contrasting groups - a) students with good fit, defined as students with an overall fit statistic of less than or equal to 1.0 and, b) students with poor fit, defined as students with an overall fit statistic greater than or equal to 1.2. The first group contained approximately 60% of the respondents with the best fit. The second group contained approximately 10% of the respondents with the worse fit. Separate functions were created for a) student background and attitude questions and, b) teacher/school questions.

Given the low correlations between fit and other variables, it was no surprise that we were unable to develop discriminant functions that accurately categorized examinees. When the discriminant function priors were set to the proportions observed in the data, everyone was classified as having good fit. When equal priors were used, only 63% of the misfits were correctly identified while a whopping 39% of the good fitting students were incorrectly labeled as misfits.

While predicted fit accuracy was low, the variables that best predicted poorly fitting examinees were quite interesting. Table 7 shows the best predictors and the percent of respondents selecting different values that were also poorly fitting examinees. Some 10.8 percent of the respondents that indicated they have never used a scientific calendar had fit statistics greater than 1.2, while 15.1 percent of the respondents that said yes had poor fit.

Some of these variables associated with poor fit are not surprising. Students that receive little instruction on measurement, students that don't think they are going to graduate, and those that feel class is often disrupted, for example, were more likely to have poor overall fit. The repeated relationship between calculator usage and fit, however, is quite interesting. Four of the variables that best discriminated between examinees with good fit and examinees

with poor fit dealt with calculator usage. It appears that students that have more calculator experience tend to have the poorest fit.

Effect of Trimming Misfitting Students on Population Proficiency

A key question in considering potential mis-fit is whether trimming misfits from the data would change any of the results. We defined misfits as those students whose overall mean square fit was greater than or equal to 1.2, computed a second set of sampling weights, and computed mean composite proficiency scores for the trimmed and untrimmed data. The fit mean square cut point of 1.2 resulting in trimming approximately 10% of the data.

Table 7
The best predictors of poor overall fit and
the percent of respondents with poor fit by item response.

Question	Percent of respondents that have fit statistics greater than 1.2		
	No	Yes	
M810401B Have you ever used a scientific calculator?	10.8	15.1	
M810501B What kind of math class are you taking?	Eighth grade 10.6	Algebra or pre-algebra 18.5	
M810103B In math class, how often do you work in small groups?	Never 10.7	Some to daily 15.9	
B007003A Do you agree: Students often disrupt class?	Agree-Strongly Disagree 12.3	Strongly agree 17.7	
M810301B How often do you use a calculator in class?	Almost always 11.0	Sometimes 16.9	Never 14.5
M810703B Do you agree: I am good in Math?	Disagree + 9.6	Undecided 12.0	Agree + 15.3
S003401A Do you expect to graduate from High School?	Yes 13.4	Don't know 20.7	No 35.8
T031601 Do you permit unrestricted use of calculators?	No 11.9	Yes 20.5	
T031504 How much emphasis is given to measurement?	Moder + 11.6	Little 17.5	None 25.0
T031201 Time students spend on math homework each day	None 3.7	15-30 minutes 12.8	45-60 minutes 22.8

Table 8 shows mean NAEP proficiencies for our trimmed and untrimmed datasets and the percent of students below behavioral anchors for the five subscales and composite scale. For all five of the subscales and for the composite, the trimmed sample mean was within one point of that for the untrimmed sample. Percents of students exceeding behavioral anchor proficiency levels were essentially identical between the two samples. Trimming the data of misfits had no impact on overall levels of proficiency. Similar patterns held when the data were analyzed by gender, race, region, parents level of education, and community type.

We should note that these values are virtually identical to the proficiency levels for the entire assessment reported by Mullis, et.al (1991). This suggests that our two percent sample is representative and that we recomputed sampling weights correctly.

Table 8
Mean Subscale and Composite Proficiencies and Anchor Level Results
for NAEP and Untrimmed and Trimmed Samples

Percents of Students at or Above Anchor Levels						
Scale	Sample	Mean Proficiency	200	250	300	350
Num & Oper	Untrimmed	265.7 (32.2)	98.5	67.7	15.9	.1
	Trimmed	265.3 (32.1)	98.5	67.7	15.5	.2
Measurement	Untrimmed	257.3 (37.8)	93.5	58.2	11.5	.2
	Trimmed	257.1 (37.9)	93.2	58.4	11.2	.3
Geometry	Untrimmed	259.2 (31.0)	97.2	61.1	10.1	0
	Trimmed	259.0 (30.9)	97.1	61.0	9.5	0
Data Analysis	Untrimmed	261.7 (37.4)	97.2	63.8	15.2	.1
	Trimmed	261.3 (37.1)	97.4	63.6	15.3	.1
Alg & Fns	Untrimmed	260.6 (33.2)	96.8	60.7	12.1	.1
	Trimmed	260.5 (33.1)	96.8	60.3	11.6	.1
Composite	Untrimmed	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	260.6 (31.3)	97.4	63.8	10.5	0

Standard deviations are in parentheses.

The above analyses shows that trimming misfitting students does not appreciably affect the mean proficiency levels or the distributions of proficiencies. Differences that exist were quite trivial; they were all less than one point on the NAEP Scale.

We repeated the same analysis using the composite NAEP scale score (computed by averaging the subscale scores) by state for untrimmed and trimmed data. As shown in Table 9, the differences were again quite small. The Spearman rank order correlation is .98.

Table 9
Trimmed and Untrimmed State Composite
Scaled Means (and Standard Deviations)

<u>State</u>	<u>Untrimmed</u>	<u>Trimmed</u>	<u>State</u>	<u>Untrimmed</u>	<u>Trimmed</u>
AL	254 (30)	255 (31)	MN	272 (27)	271 (27)
AR	256 (34)	255 (34)	MT	274 (27)	271 (26)
AZ	258 (29)	258 (29)	NC	256 (31)	257 (31)
CA	261 (33)	261 (34)	ND	276 (29)	276 (29)
CO	264 (27)	261 (28)	NE	274 (25)	274 (26)
CT	266 (36)	266 (38)	NH	268 (24)	268 (23)
DC	236 (30)	236 (30)	NJ	273 (30)	273 (30)
DE	263 (35)	268 (34)	NM	250 (23)	250 (24)
FL	263 (37)	263 (37)	NY	265 (36)	264 (35)
GA	248 (30)	248 (30)	OH	274 (27)	273 (27)
GU	244 (35)	242 (34)	OK	262 (26)	260 (25)
HI	249 (37)	250 (38)	OR	265 (31)	267 (31)
IA	276 (27)	275 (28)	PA	264 (28)	267 (23)
ID	267 (25)	267 (24)	RI	251 (33)	248 (32)
IL	251 (32)	254 (32)	TX	255 (27)	254 (27)
IN	272 (27)	269 (25)	VA	266 (31)	262 (32)
KY	249 (27)	250 (28)	VI	241 (35)	241 (36)
LA	245 (25)	244 (26)	WI	271 (24)	271 (25)
MD	257 (34)	256 (33)	WV	255 (27)	253 (26)
MI	268 (26)	269 (25)	WY	268 (22)	266 (22)

Chapter 4: Full 1990 & 1992 Assessments

Methods

This methods section describes the data from the full 1990 and 1992 assessment data that we used and how we computed fit.

Data

In the winters of 1990 and 1992, NAEP also conducted regular assessments of students in fourth, eighth and twelve grade using large nationally representative samples. We analyzed the full grade 8 datasets for the 1990 mathematics, 1990 reading, 1992 estimation, 1992 mathematics, and 1992 reading assessments contained in the Secondary-Use Data Files. Each student's record consisted of responses to blocks of items, common background items, subject area background items, the students' teacher's responses to a questionnaire, and the students' school administrator's responses to a questionnaire. Each record also contained plausible value ability estimates for each subscale and a composite.

The mathematics and reading assessments contained items that involved constructed responses and the use of raters. The mathematics assessments also contained large numbers of items that involved the use of calculators. The estimation assessment contained neither.

Interim files containing the item parameters, correct responses, scale, and item type (rater, no rater, calculator, no calculator) were created from the SPSS control cards and the data layout files provided by ETS. The data in the interim file and the theta values and item response patterns from the full dataset were then analyzed to provide fit statistics over all attempted items, over items that had a rater, over those that did not, over items that involved calculator use and over items that did not. The data were inserted in a copy of the full

dataset and then analyzed using SPSS. The software is appended to this report. Because the formats are similar for each datafile, the same programs were used with each dataset.

Computation of fit

As with the analysis of the TSA data, we needed estimates of the probability of a correct response, P_{ij} , and the observed response u_{ij} , for each individual. P_{ij} was computed using the 3 parameter logistic model, estimates for subscale θ 's that appear in the public use dataset, and item parameters provided by ETS. Values for u_{ij} were taken from the public use dataset. Subscale thetas were used rather than the composite theta. As with the ETS analysis, omits were coded as incorrect responses.

Results

We examined

- 1) overall fit,
- 2) whether there are differences in fit across rater versus no rater,
- 3) whether there are differences in fit across constructed response versus multiple choice items,
- 4) whether there are differences in fit across calculator use versus no calculator use,
- 5) the correlation between fit and total score, and
- 6) the effects of trimming mis-fitting students.

Overall Fit

The distributions of the mean square fit statistic for five assessments are shown in Table 10. With the exception of both the 1990 and 1992 national mathematics assessments, fit was quite good. The mean fit statistics approached 1.0 with little variance. The mathematics assessments had a great deal of mis-fit with large variances.

Table 10
Distribution of the mean square fit
statistic for five National Assessments

<u>Assessment</u>	<u>Mean</u>	<u>s.d.</u>	<u>Weighted N</u>
1990 Mathematics	1.26	.41	8,634
1990 Reading	1.01	.28	8,708
1992 Mathematics	1.17	.32	10,291
1992 Reading	1.11	.36	12,630
1992 Estimation	1.06	.18	2,416

Rater versus no Rater

To compare the fit on items that involved raters versus items that did not, we divided each examinee's response vector into two groups, one for each type of item. As shown in Table 11, there were statistically significant differences in all cases. Except for the 1992 reading assessment, mean fit was better on items that involved raters. Items that involved raters, however, showed a greater variance than items that did not. In all cases, the differences are small.

Table 11
Mean fit on rater versus no rater items

	Rater Mean	No Rater Mean	Rater s.d.	No Rater s.d.	t	df	p
1990 Mathematics							
	1.0855	1.3103	.500	.509	-32.15	8632	< .001
1990 Reading							
	.9713	1.0451	.787	.359	-5.66	3720	< .001
1992 Mathematics							
	.9709	1.2067	.613	.373	-33.43	10290	< .001
1992 Reading							
	1.2288	1.0949	.622	.505	18.55	10303	< .001

Constructed Response versus multiple choice questions

To compare the fit on constructed response versus multiple choice questions, we divided each examinee's response vector into two groups, one for each type of item. As shown in Table 12, there were statistically significant differences in all cases. For mathematics tests, the best fit was on multiple choice items; for reading it was on constructed response items. The differences, however, are slight. The effect sizes are minimal. The results are similar to that of rater versus no rater as the categories are highly related. That is, many of the constructed response items involved raters and all of the multiple choice items did not involve raters.

Table 12
Mean fit on constructed response versus multiple choice items

	Constr Rspnse Mean	Multiple Choice Mean	Constr Rspnse s.d.	Multiple Choice s.d.	t	df	p
1990 Mathematics							
	1.3183	1.2655	.613	.479	7.84	8629	.000
1990 Reading							
	.9713	1.0451	.787	.359	-5.66	3720	.000
1992 Mathematics							
	1.2032	1.1775	.612	.383	4.12	10283	.000
1992 Reading							
	1.0922	1.1273	.450	.632	-5.57	12626	.000

Calculator Use versus no Calculator

One innovative item type introduced in the 1990 assessment was the use of calculators. As shown in Table 13, fit was quite good on the items that did not involve a calculator and quite bad on items that did. Further the variance was much greater on the calculator use items.

Table 13
Mean Fit on Items with and without Calculator Use

	Calc Mean	No Calc Mean	Calc s.d.	No Calc s.d.	t	df	p
1990 Mathematics							
	2.0054	1.0677	1.022	.254	73.57	6116	.000
1992 Mathematics							
	2.2376	.9831	.834	.215	112.57	5873	.000

Correlations of fit statistics with total scaled score.

Up to seven fit statistics were computed for each examinee - one overall attempted items, and one for each item type. If the assessments were equally accurate across examinee ability, then the correlations should approach zero. If examinees respond in the same way to various item types, then the correlation of fit statistics with total score would be similar across item types.

As shown on table 14, the correlations between overall fit and total score approach zero for the 1992 mathematics and 1992 estimation assessments. The correlations also approach zero for the constructed response items in the 1990 mathematics assessments. There are moderate correlations with reading ability and fit. Higher ability examinees tended to have poor fit. The correlations were different across item types for all the assessments except the 1992 reading assessment.

There were surprisingly large correlations between fit and total score for the 1990 mathematics assessment for calculator use items, no rater items, and multiple choice items; and between fit and total score for the 1992 Reading items. These sets are not independent. As we will show later, the use of calculators is the significant variable in the mathematics

correlations. It appears that the correlations for reading are due to constructed response items.

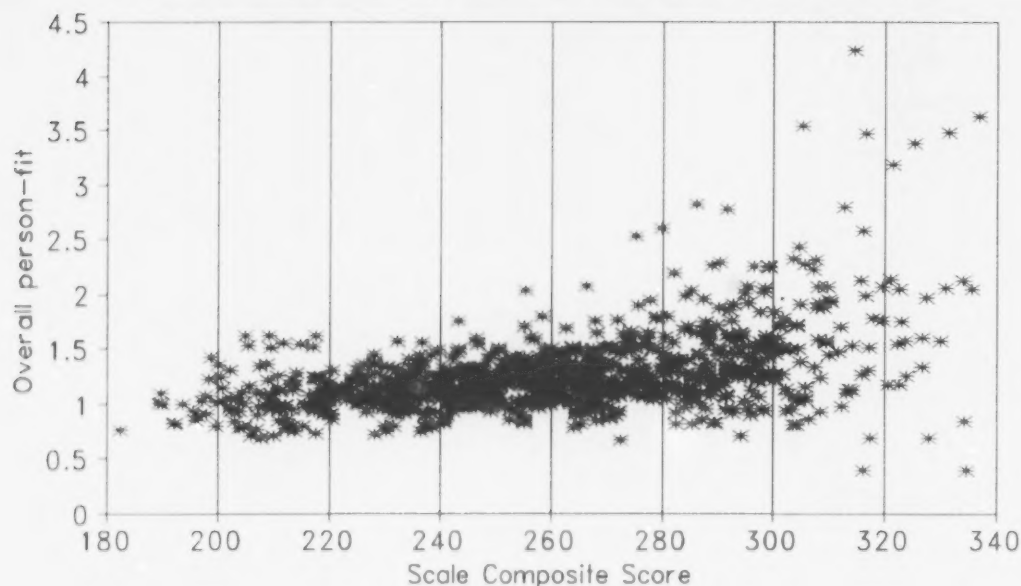
These large linear correlations are quite disturbing. The situation is worse when one considers that the variance in fit is not homogeneous across measured mathematics ability. As shown in Figure 3, the variance in fit is much greater for high ability students.

Table 14
Correlations (N) of various fit statistics and Scaled Score

	Total Scaled Score				
	1990 Mathe- matics	1992 Mathe- matics	1992 Estima- tion	1990 Reading	1992 Reading
Over all	.4301 (8634)	.0332 (10291)	.0088 (2416)	.1615 (8708)	.2710 (12630)
Rater	.1147 (8634)	.1378 (10291)		.2535 (3723)	.1793 (10304)
No Rater	.4359 (8633)	.0096 (10291)		.1479 (8708)	.1772 (12630)
Calculator	.5631 (6117)	-.2827 (5874)			
No Calculator	.1647 (8634)	.1713 (10291)			
Constructed Response	.0678 (8631)	-.0973 (10284)		.2535 (3723)	.2961 (12630)
Multiple Choice	.4604 (8633)	.0307 (10291)		.1479 (8708)	.1213 (12627)

Figure 3

Overall Fit and total score (1990 Mathematics)



Item Fit

A key question is whether the poor fit on the calculator use items is due to differences in the way examinees approach these items or in the accuracy of the item calibration. Items that involved raters and items that involved constructed responses have calculated item response theory item parameters. Items that involved calculator use were analyzed above, and by the NAEP team, using the same item parameters as the identical item when administered without a calculator.

To address the question of item calibration accuracy, we computed item fit statistics by summing across examinees rather than across items. As shown on Table 15, the item fit was much worse for items for which the students used calculators. Calculator use items are

defined here as those items coded as calculator use items in the NAEP Dataset. Since the calculators were not removed when the students responded to non-calculator items, the "no calculator" items probably contain some items for which the students actually did use calculators.

Table 15
Mean Fit (and standard deviation) by item type

	Rater	No Rater	Constructed Response	Multiple Choice	All Items
Calculator Use		2.4238 (2.3561)	2.5847 (2.0441)	2.3721 (2.4803)	2.4238 (2.3561)
No Calculator	1.1256 (.3467)	1.1167 (.3396)	1.1256 (.3467)	1.1167 (.3396)	1.1190 (.3402)

Do misfitting students alter the results?

The final results question is whether trimming misfitting students would have made a difference in mean aggregate scores. The appendix contains tables showing trimmed and untrimmed means, standard deviation and confidence intervals for each of the key NAEP reporting groups - total, race, gender, region, type of community and size of community. From Table 16, which summarizes the findings, one can see that the difference in the means from trimmed versus untrimmed groups is rather small. Positive values indicate that the untrimmed mean score was higher than the mean score when misfitting students were trimmed from the dataset. An asterisk indicates statistical significance. Since the typical variance in NAEP Scales scores is about 35, most of these differences represent a very small effect, even if they may be statistically significant.

The biggest differences are for the 1990 Mathematics and 1992 Reading Assessments. Trimming misfitting students tends to lower average scores by 10 and 5 points respectively. Trimming had its greatest effect on the average scores for Asian students' scores.

Table 16
NAEP Scales Score Mean Differences for untrimmed versus trimmed groups
for 5 National Eighth Grade Assessments.

Group	1990 Math	1990 Reading	1992 Estmntn	1992 Math	1992 Reading
Total	9.44*	2.80*	0.34	0.28	4.77*
White	8.03*	2.65*	2.01	0.43	3.49*
Black	6.68*	2.57	-3.08	-1.00	5.76*
Hispanic	7.82*	2.24	-2.55	-2.16	5.71*
Asian	16.55*	1.77	-0.06	10.08	6.86
Male	9.79*	2.48*	1.13	-0.82	4.33*
Female	9.02*	2.86*	-0.58	1.53	4.57*
Northeast	9.39*	2.56	-1.68	0.41	4.79
Southeast	8.31*	2.64	-0.02	-0.60	4.67*
Central	9.02*	2.24	1.45	0.13	3.56*
West	9.98*	3.48*	1.35	0.86	5.98*
Ext Rural	5.60*	2.41	1.71		2.87
Dis urban	9.27*	2.08	-5.23		5.41*
Adv Urban	9.41*	3.14	3.36		4.27*
Other	9.13*	2.78*	0.65		4.39*
Ext Rural	5.60*	2.41	1.71	-0.74	2.87
Low Metro	9.27*	2.08	-5.23	-1.85	5.41*
High Metro	9.41*	3.14	3.36	-0.01	4.27*
Big City	10.96*	4.02	-0.34	-0.37	5.15*
Urban Fringe	9.85*	2.16	2.42	1.92	5.64*
Medium City	10.37*	3.43	0.42	1.04	4.18*
Small place	7.83*	2.47	0.60	-0.30	-3.56

*=statistically significant at the .05 level

Chapter 5: Discussion

Given the length of testing time, the complexity of the assessment, and the politically charged atmosphere surrounding the NAEP Trial State Assessment, the primary author expected to find a fair number of students with atypical response patterns. Contrary to this expectation, fit was exceptionally good. The average fit was less than 1.0 and very few examinees had extreme values. There was also exceptionally good fit when the data was examined by state, by block order, and by traditional NAEP reporting variables.

We also liberally trimmed the dataset to eliminate the students with the worst fit. We found no meaningful differences in the mean NAEP Scale scores between the trimmed and untrimmed data. An analysis of state level data had the same conclusion; there is no meaningful difference in state composite scores for trimmed and untrimmed data.

The 1990 NAEP Assessment was composed mostly of traditional multiple choice items. We noted that the worse fit in the assessment was for the block and scale using open-ended questions. We also noted that the students with the most experience using calculators tended to be more highly represented in the group of students classified as having poor fit.

Given the current interest in alternative assessment and the increase in the use of non-traditional item types in large scale assessment, we replicated our analysis using the responses made by grade 8 students to the whole 1990 mathematics, 1990 reading, 1992 estimation, 1992 reading, and 1992 mathematics assessments. We specifically analyzed the data by rater versus no rater, calculator versus no calculator, and constructed response versus multiple choice questions.

Again fit was exceptionally good, especially for the 1990 reading, 1992 estimation, and 1992 mathematics assessments. The average fit approached 1.0, differences between rater versus no rater were negligible as were the differences between constructed response and

multiple choice items. Correlations between total score and degree of mis-fit were around 0.00, and trimmed misfitting students tended not to make any difference.

There were, however, several notable exceptions. Fit was poor for the overall 1990 mathematics assessment and extremely bad for items that involved calculator usage. Further there was a large positive correlation between degree of mis-fit and total scaled score for both the 1990 mathematics and the 1992 reading assessments. With these sizeable correlations, trimming misfitting students lowered average test scores.

Thus, for the trial state assessment and three of the total NAEP Assessments, fit was exceptional. Two other assessments had sizeable numbers of students whose response patterns did not fit the model. The scores of these students may have seriously skewed the results upward. Later we show that there may be a problem with the measurement model.

There are three necessary and sufficient conditions for good person-fit:

- 1) ability must be accurately estimated,
- 2) the item parameters must be accurately estimated, and
- 3) individuals must respond in a consistent manner.

In the case of the trial state assessment and the 1992 estimation assessments fit was exceptionally good and all three conditions were met. These results strongly suggest that individuals can be consistent in their response patterns and that, when scores are aggregated, person fit statistics have little to offer.

The poorer fit on the whole 1990 mathematics assessment appears to be due to poor item fit rather than response inconsistencies. The item fit for items when administered with a calculator was much worse than the fit for the same items when administered without a calculator. The item characteristic curves do not fit the data equally well under both conditions. Most likely, different sets of item parameters are needed.

We noted earlier that there are sizable positive correlations between total ability and degree of misfit for the 1990 Mathematics and 1992 reading assessments and that fit is worse and more variable for students with high ability. If we assume that individuals tend to respond consistently (since they do for other assessments and since low ability students are consistent) and that the item parameters are sufficiently accurate (since they appear so for middle and low ability students), then one might think that the problem lies in the estimation of ability. However, we found better fit for students with high percentages of correct items and fewer reached items on these assessments. Thus, the imputation procedure appeared to be more accurate in these cases rather than less accurate. Further exploration is needed to resolve this anomaly. This might indicate a problem in the penalty for guessing, particularly when the student has partial knowledge¹.

While we conclude that this person fit statistic has little to offer in the analysis and reporting of NAEP data, we do not generalize to other assessments. More resources go into the development and analysis of NAEP than any other assessment program. As a result, the items have always been of the highest quality. Further, the interest with NAEP is in aggregated scores, not individual scores. In testing situations where individual scores make a difference, person fit statistics can be used to identify individuals with strange response patterns. A reexamination of their scores or reassessing their abilities could well lead to different decisions concerning that individual.

¹ This possibility was pointed out by Mike Cohen in reviewing an earlier version of this manuscript.

Appendix 1: Subscale Tables

Table A-1
Correlations of Model Fit Statistics Between Subscales and Item Blocks
on the 1990 Trial State Assessment

	ME	GE	DA	AL	M3	M4	M5	M6	M7	M8	M9
NO	.11	.08	.01	.11	.44	.35	.14	.35	.22	.49	.49
ME		.04	.02	.01	.24	.37	.09	.40	.18	.29	.02
GE			.01	.02	.23	.34	.35	.30	.38	.09	.18
DA				.05	.32	.20	.40	.13	.19	.17	.18
AL					.20	.23	.20	.22	.33	.27	.42
M3						.15	-.06	.06	-.02	.03	.05
M4							.05	.06	.06	-.08	.06
M5								-.03	.04	-.20	.09
M6									.05	-.12	-.07
M7										.17	-.12
M8											-.03

Table A-2
Correlations Between Fit Statistics and Plausible Values
on the 1990 Trial State Assessment

Plausible Value	Num & Op	Measmt	Geometry	Data Anal	Alg & Fns
Num & Oper	-.02	-.02	-.04	-.07	-.01
Measurement	.00	-.08	-.04	-.06	-.02
Geometry	-.02	.02	-.07	-.04	-.04
Data Anal	.00	.01	-.03	-.08	-.01
Alg & Fns	.00	.01	-.04	-.06	-.01

Table A-3
Correlations Greater than .10 or Less than -.10
Between Subscale Fit Statistics and Background Variables
on the 1990 Trial State Assessment

Background Variable	N & O	Meas.	Geom.	D A	A & F
> 25 books at home?					-.10
Family get mag?				.11	
Do text probs?				.15	
Take math tests?			-.10		
Used scient. calc?			-.11		
Use math on job?				-.11	
Tchr cert/jr hi math?			-.12		.11
Tchr acad. degree	.11				
Tchr undergrad major			-.12		
No tchr under. study		.11			
Tchr grad math major		.11			
Tchr courses/math	.10				
Tchr courses/alg					.12
Tchr courses/geom					.13
Tchr courses/comp sci		.12			
Class ability			-.11		
Hrs. in math/week					-.11
Time math homework	.11				
Work in small groups?		-.11			
Often use calculators?				.10	
Take tchr math tests?					.12
Student reports?					.11
Emphasis: ratios		.10		.12	
Emphasis: percents				.12	
Emphasis: geometry				.14	
Emphasis: algebra				.12	
Emphasis: ideas		-.11			
Permit calculators	-.12	-.14		.12	-.12
Calc on tests?				.12	
Access/school calc?	.11				.13
Computer use for math		.11			

Table A-4
Results of Discriminant Analyses by Subscale
on the 1990 Trial State Assessment

Numbers and Operations Subscale

Type of Variables	No. Vars.	No. Valid Cases	Prop. Fit	Prop. Misfit	Percent Correct Class.	Mult. R	Mult. R-sq.
Student/School	43	805	.94	.06	95%	.25	.06
Math Background	22	1,266	.94	.06	94	.18	.03
Teacher Backgrd.	74	434	.95	.05	98	.54	.29
Combination	31	1,089	.94	.06	95	.29	.09

Measurement Subscale

Type of Variables	No. Vars.	No. Valid Cases	Prop. Fit	Prop. Misfit	Percent Correct Class.	Mult. R	Mult. R-sq.
Student/School	43	805	.85	.15	85%	.26	.07
Math Background	22	1,266	.85	.15	85	.16	.03
Teacher Backgrd.	74	434	.86	.14	89	.51	.26
Combination	38	912	.85	.15	85	.31	.09

Table A-4 (continued)

Geometry Subscale

Type of Variables	No. Vars.	No. Valid Cases	Prop. Fit	Prop. Misfit	Percent Correct Class.	Mult. R	Mult. R-sq.
Student/School	43	805	.94	.06	95†	.25	.06
Math Background	22	1,266	.94	.06	94	.18	.03
Teacher Backgrd.	74	434	.95	.05	98	.54	.29
Combination	31	1,089	.94	.06	95	.29	.09

Data Analysis, Statistics, & Probability Subscale

Type of Variables	No. Vars.	No. Valid Cases	Prop. Fit	Prop. Misfit	Percent Correct Class.	Mult. R	Mult. R-sq.
Student/School	43	805	.73	.27	76†	.30	.09
Math Background	22	1,266	.70	.30	71	.26	.07
Teacher Backgrd.	74	434	.71	.29	78	.44	.19
Combination	35	887	.72	.28	74	.31	.09

BEST COPY AVAILABLE

Table A-4 (continued)

Algebra and Functions

Type of Variables	No. Vars.	No. Valid Cases	Prop. Fit	Prop. Misfit	Percent Correct Class.	Mult. R	Mult. R-sq.
Student/School	43	805	.86	.14	87%	.33	.11
Math Background	22	1,266	.89	.11	89	.15	.02
Teacher Backgrd.	74	434	.88	.12	91	.49	.24
Combination	45	836	.90	.10	90	.30	.09

Table A-5
Means and Standard Deviations of Fit Statistics
by Sex, Race and Subscale
on the 1990 Trial State Assessment

Numbers and Operations

Race/Ethnicity	Male	Females	Total Race/Eth.
White	.99 (.29)	.92 (.31)	.96 (.30)
Black	.97 (.31)	.93 (.28)	.99 (.25)
Hispanic	1.00 (.25)	.98 (.25)	.99 (.25)
Asian	.95 (.24)	.98 (.26)	.96 (.28)
Amer. Indian	.94 (.29)	.93 (.24)	.94 (.24)
Total Sex	.98 (.29)	.93 ()	.96 (.29)

Standard deviations of mean fit statistics are in parentheses.

Measurement

Race/Ethnicity	Males	Females	Total Race/Eth.
White	.97 (.46)	.97 (.39)	.97 (.43)
Black	.91 (.32)	.96 (.32)	.94 (.32)
Hispanic	.88 (.32)*	.97 (.38)	.93 (.36)
Asian	.93 (.44)	.91 (.40)	.92 (.42)
Amer. Indian	.94 (.30)	.83 (.33)*	.89 (.32)
Total Sex	.95 (.43)	.96 (.38)	.96 (.41)

Standard deviations of mean fit statistics are in parentheses.

Geometry

Race/Ethnicity	Males	Females	Total Race/Eth.
White	.92 (.31)	.90 (.30)	.91 (.31)
Black	.96 (.26)	.89 (.32)	.93 (.29)
Hispanic	.91 (.27)	.88 (.29)	.89 (.28)
Asian	.92 (.32)	.92 (.27)	.92 (.29)
Amer. Indian	.99 (.22)*	.96 (.33)	.98 (.27)
Total Sex	.92 (.30)	.90 (.30)	.91 (.30)

Standard deviations of mean fit statistics are in parentheses.

Data Analysis, Statistics, & Probability

Race/Ethnicity	Males	Females	Total Race/Eth.
White	1.11 (.57)	1.14 (.57)	1.12 (.57)
Black	1.17 (.63)	1.22 (.73)	1.19 (.68)
Hispanic	1.22 (.57)	1.15 (.53)	1.18 (.55)
Asian	1.21 (.44)	1.16 (.60)	1.19 (.52)
Amer. Indian	1.10 (.47)	.95 (.43)*	1.04 (.46)
Total Sex	1.13 (.57)	1.15 (.59)	1.14 (.58)

Standard deviations of mean fit statistics are in parentheses.

Algebra and Functions

Race/Ethnicity	Males	Females	Total Race/Eth.
White	.94 (.42)	.92 (.40)	.93 (.41)
Black	.90 (.38)	.90 (.38)	.90 (.38)
Hispanic	.90 (.40)	.97 (.41)	.93 (.41)
Asian	.94 (.42)	.82 (.34)*	.88 (.38)
Amer. Indian	.89 (.41)	.94 (.40)	.91 (.40)
Total Sex	.93 (.41)	.92 (.39)	.93 (.40)

Standard deviations of mean fit statistics are in parentheses.

Overall

Race/Ethnicity	Males	Females	Total Race/Eth.
White	.98 (.19)	.96 (.16)	.97 (.18)
Black	.97 (.14)	.96 (.17)	.97 (.15)
Hispanic	.98 (.15)	.98 (.15)	.98 (.15)
Asian	.99 (.18)	.95 (.17)*	.97 (.18)
Amer. Indian	.97 (.16)	.92 (.17)	.95 (.17)
Total Sex	.98 (.18)	.96 (.16)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

Table A-6
Means and Standard Deviations of Fit Statistics
by Select Demographic variables and Subscale
on the 1990 Trial State Assessment

by Region

Region	Num & Oper	Measurement	Geometry	Data Anal.	Alg & Fns	Overall Fit
Northeast	.98 (.30)	.96 (.44)	.93 (.32)	1.10 (.54)	.97 (.44)	.98 (.19)
Southeast	.93 (.29)	.94 (.36)	.90 (.28)	1.20 (.64)	.89 (.41)	.96 (.17)
Central	.96 (.30)	.98 (.43)	.92 (.30)	1.06 (.55)	.94 (.39)	.97 (.17)
West	.97 (.27)	.95 (.40)	.90 (.30)	1.18 (.55)	.92 (.38)	.97 (.16)
Nation	.96 (.29)	.96 (.41)	.91 (.30)	1.14 (.58)	.93 (.41)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

by Type of Community

Community	Num & Oper	Measurement	Geometry	Data Anal.	Alg & Fns	Overall Fit
Extreme rural	.92 (.28)	.95 (.37)	.91 (.30)	1.17 (.58)	.91 (.37)	.95 (.16)
Adv. Urban	.94 (.26)	.91 (.34)	.90 (.29)	1.14 (.64)	.93 (.38)	.95 (.16)
Disadv. Urban	.98 (.31)	.95 (.41)	.88 (.32)	1.09 (.58)	.95 (.45)	.98 (.19)
Other	.97 (.29)	.98 (.43)	.91 (.30)	1.13 (.57)	.92 (.41)	.97 (.18)
Nation	.96 (.29)	.96 (.41)	.91 (.30)	1.14 (.58)	.93 (.41)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

by Parents' Highest Level of Education

Education	Num & Oper	Measurement	Geometry	Data Anal.	Alg & Fns	Overall Fit
Did Not Finish High School	.93 (.24)	.94 (.33)	.90 (.28)	1.24 (.56)	.95 (.43)	.97 (.15)
Grad. from High School	.97 (.27)	.94 (.36)	.92 (.30)	1.17 (.61)	.97 (.42)	.98 (.16)
Some Educ. after High School	.94 (.29)	1.00 (.46)	.93 (.29)	1.16 (.57)	.90 (.39)	.97 (.18)
Grad. from College	.97 (.31)	.96 (.43)	.90 (.31)	1.09 (.55)	.92 (.40)	.96 (.18)
Nation	.96 (.29)	.96 (.41)	.91 (.30)	1.14 (.58)	.93 (.41)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

by Students' Perceptions of Mathematics

Attitude	Num & Oper	Measurement	Geometry	Data Anal.	Alg & Fns	Overall Fit
Strongly positive	.96 (.33)	.93 (.43)	.90 (.32)	1.08 (.57)	.91 (.40)	.96 (.18)
Positive	.97 (.27)	.96 (.41)	.91 (.30)	1.17 (.58)	.94 (.41)	.98 (.17)
Undecided or Negative	.94 (.26)	.99 (.38)	.93 (.29)	1.15 (.56)	.93 (.39)	.97 (.16)
Nation	.96 (.29)	.96 (.41)	.91 (.30)	1.14 (.58)	.93 (.41)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

by Appropriateness of Calculator Usage

Appropriateness	Num & Oper	Measurement	Geometry	Data Anal.	Alg & Fns	Overall Fit
High	.95 (.28)	.94 (.40)	.89 (.29)	1.22 (.59)	.93 (.44)	.97 (.17)
Other	.97 (.27)	.99 (.42)	.92 (.31)	1.10 (.58)	.94 (.42)	.98 (.17)
Nation	.96 (.29)	.96 (.41)	.91 (.30)	1.14 (.58)	.93 (.41)	.97 (.17)

Standard deviations of mean fit statistics are in parentheses.

Table A-7
Mean Composite Proficiency and Anchor Level Results
for NAEP and Untrimmed and Trimmed Samples
on the 1990 Trial State Assessment
by Gender

Percents of Students at or Above Anchor Levels							
Sex	Sample	Pct. of Students	Mean Proficiency	200	250	300	350
Males	NAEP	51	262.	97	64	14	0
	Untrimmed	53	264.3 (31.0)	98.5	66.9	13.2	0
	Trimmed	53	263.8 (31.0)	98.4	66.4	13.0	0
Females	NAEP	49	260.	97	64	10	0
	Untrimmed	47	257.0 (31.4)	96.1	60.7	8.2	0
	Trimmed	47	256.9 (31.3)	96.1	61.0	7.9	0
Total	NAEP	100	261.	97	64	12	0
	Untrimmed	100	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	100	260.9 (31.3)	97.4	63.8	10.5	0

Standard deviations of proficiencies are in parentheses.

by Race/Ethnicity

Percents of Students at or Above Anchor Levels

Race/ethnicity	Sample	Pct. of Students	Mean Proficiency	200	250	300	350
White	NAEP	70	269	99	74	15	0
	Untrimmed	59	270.5 (27.0)	99.5	76.8	14.2	0
	Trimmed	69	270.0 (26.8)	99.5	76.7	13.7	0
Black	NAEP	16	236	89	30	2	0
	Untrimmed	15	232.5 (25.5)	92.2	27.0	.6	0
	Trimmed	15	233.0 (25.9)	92.0	28.1	.6	0
Hispanic	NAEP	10	243	93	41	3	0
	Untrimmed	8	236.4 (26.4)	92.0	29.7	.6	0
	Trimmed	8	235.7 (26.5)	92.1	27.2	.7	0
Asian	NAEP	2	280*	97	80	31	1
	Untrimmed	5	257.7 (39.5)	92.2	59.8	16.7	0
	Trimmed	5	257.3 (40.5)	91.8	57.6	17.8	0
Amer. Indian	NAEP	2	246*	97	45	1	0
	Untrimmed	2	250.1 (23.5)	97.0	51.3	4.4	0
	Trimmed	2	250.3 (24.3)	96.9	51.8	4.8	0
Total	NAEP	100	261	97	64	12	0
	Untrimmed	100	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	100	260.9 (31.3)	97.4	63.8	10.5	0

Standard deviations of proficiencies are in parentheses.

* Variability of this statistic cannot be determined (Mullis et al., 1991)

by Geographic Region

Percents of Students at or Above Anchor Levels

Region	Sample	Pct. of Students	Mean Proficiency	200	250	300	350
Northeast	NAEP	21	269	99	72	16	0
	Untrimmed	23	261.8 (32.8)	96.5	64.3	11.6	0
	Trimmed	23	261.7 (32.0)	96.9	65.5	10.5	0
Southeast	NAEP	24	253	94	52	8	0
	Untrimmed	26	253.3 (31.5)	96.4	52.3	9.3	0
	Trimmed	26	253.2 (32.0)	96.1	52.4	9.7	0
Central	NAEP	25	265	98	70	12	0
	Untrimmed	23	270.9 (28.3)	99.3	77.3	15.5	0
	Trimmed	23	270.3 (27.8)	99.2	76.8	15.1	0
West	NAEP	30	261	97	63	12	0
	Untrimmed	26	261.3 (28.8)	98.5	66.5	8.1	0
	Trimmed	26	260.9 (29.0)	98.3	65.5	8.1	0
Total	NAEP	100	261	97	64	12	0
	Untrimmed	100	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	100	260.9 (31.3)	97.4	63.8	10.5	0

Standard deviations of proficiencies are in parentheses.

by Parents' Highest Level of Education

Percents of Students at or Above Anchor Levels

Education	Sample	Pct. of Students	Mean Proficiency	200	250	300	350
Did Not Finish	NAEP	10	243	96	37	1	0
High School	Untrimmed	8	244.3 (23.4)	97.9	35.2	1.7	0
	Trimmed	8	244.6 (23.7)	97.9	35.1	1.8	0
Graduated High School	NAEP	25	254	97	56	5	0
	Untrimmed	26	251.5 (27.8)	96.8	55.5	3.3	0
	Trimmed	26	251.4 (27.3)	96.7	55.6	3.3	0
Some Education after	NAEP	17	266	99	71	12	0
High School	Untrimmed	21	267.5 (27.1)	99.7	75.3	11.7	0
	Trimmed	20	266.6 (26.9)	99.7	74.7	11.3	0
Graduated College	NAEP	39	274	99	78	21	0
	Untrimmed	37	273.2 (30.5)	98.6	76.9	19.8	0
	Trimmed	37	272.8 (30.8)	98.6	76.8	19.5	0
Total	NAEP	100	261	97	64	12	0
	Untrimmed	100	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	100	260.6 (31.3)	97.4	63.8	10.5	0

Standard deviations of proficiencies are in parentheses.

by Type of Community

Percents of Students at or Above Anchor Levels

Type of Community	Sample	Pct. of Students	Mean Proficiency	200	250	300	350
Advantaged Urban	NAEP	10	281*	100	83	26	1
	Untrimmed	17	261.3 (30.8)	97.5	63.3	11.7	0
	Trimmed	17	261.5 (31.1)	97.4	62.8	12.6	0
Disadvantaged Urban	NAEP	10	249*	95	48	7	0
	Untrimmed	10	235.0 (29.5)	89.5	32.1	1.3	0
	Trimmed	10	235.3 (29.4)	90.3	32.9	1.5	1
Extreme Rural	NAEP	10	256*	97	56	6	0
	Untrimmed	10	279.4 (29.1)	100	83.7	25.6	0
	Trimmed	10	277.9 (29.6)	100	81.6	24.2	0
Other	NAEP	70	261	97	64	12	0
	Untrimmed	56	261.8 (29.5)	98.3	66.2	9.8	0
	Trimmed	56	261.1 (29.4)	98.1	65.9	9.3	0
Total	NAEP	100	261	97	64	12	0
	Untrimmed	100	260.9 (31.3)	97.2	63.9	10.9	0
	Trimmed	100	260.9 (31.3)	97.4	63.8	10.5	0

Standard deviations of proficiencies are in parentheses.

* Variability of this statistic cannot be determined (Mullis et al., 1991)

Table A-8
Trimmed and Untrimmed Mean scores for the
1990 NAEP Mathematics Assessment

		Mean	Standard Deviation	95% Confidence Interval		
TOTAL						
	untrimmed	258.5172	32.2279	257.8373	To	259.1970*
	trimmed	249.0799	30.8546	248.1808	To	249.9790
RACE						
WHITE-	untrimmed	266.8627	29.1473	266.1230	To	267.6023*
	trimmed	258.8332	28.0849	257.8181	To	259.8483
BLACK-	untrimmed	233.1709	28.0055	231.6691	To	234.6726*
	trimmed	225.3482	25.4299	223.5954	To	227.1011
HISPANIC-	untrimmed	239.7671	28.0275	237.9992	To	241.5349*
	trimmed	233.0857	25.7866	231.0142	To	235.1573
ASIAN-	untrimmed	278.5961	34.5224	273.8374	To	283.3549*
	trimmed	262.0466	33.9119	254.4543	To	269.6388
Gender						
MALE-	untrimmed	258.3403	32.9707	257.3752	To	259.3054*
	trimmed	248.5487	30.9012	247.3141	To	249.7834
FEMALE-	untrimmed	258.7084	31.4079	257.7524	To	259.6644*
	trimmed	249.6841	30.7977	248.3716	To	250.9966
REGION						
NORTHEAST-	untrimmed	264.7214	31.1570	263.2355	To	266.2073*
	trimmed	255.3349	30.2086	253.2927	To	257.3771
SOUTHEAST-	untrimmed	248.3452	31.3465	246.9935	To	249.6969*
	trimmed	240.0377	29.4167	238.3580	To	241.7175
CENTRAL-	untrimmed	263.6396	30.3975	262.3713	To	264.9079*
	trimmed	254.6180	29.4340	252.8827	To	256.3532
WEST-	untrimmed	258.5333	32.6123	257.2756	To	259.7910*
	trimmed	248.5546	31.6491	246.8924	To	250.2168

		Standard			
	Mean	Deviation	95% Confidence Interval		
Type of Community					
EXTREME RURAL- untrimme	255.0790	30.0532	253.2485	To	256.9095*
trimmed	249.4760	29.3623	247.0814	To	251.8705
DISADVANTAGED URBAN- un	244.4610	31.2435	242.4327	To	246.4893*
trimmed	235.1872	27.8263	232.8632	To	237.5113
ADVANTAGED URBAN- untri	279.0590	28.8109	277.1615	To	280.9565*
trimmed	269.6470	29.4912	266.6074	To	272.6865
OTHER (NON-EXTREME)- un	258.2018	31.7353	257.3845	To	259.0191*
trimmed	249.0691	30.4346	247.9849	To	250.1534
Size of Community					
EXTREME RURAL- untrimme	255.0790	30.0532	253.2485	To	256.9095*
trimmed	249.4760	29.3623	247.0814	To	251.8705
LOW METROPOLITAN- untri	244.4610	31.2435	242.4327	To	246.4893*
trimmed	235.1872	27.8263	232.8632	To	237.5113
HIGH METROPOLITAN- untr	279.0590	28.8109	277.1615	To	280.9565*
trimmed	269.6470	29.4912	266.6074	To	272.6865
MAIN BIG CITY- untrimme	252.7263	32.9308	250.4083	To	255.0444*
trimmed	241.7657	29.8807	238.9340	To	244.5974
URBAN FRINGE- untrimmed	264.5645	32.8053	262.3758	To	266.7533*
trimmed	254.7172	32.8063	251.5482	To	257.8863
MEDIUM CITY- untrimmed	257.9364	31.3972	256.0518	To	259.8211*
trimmed	247.5675	29.5488	245.0799	To	250.0551
SMALL PLACE- untrimmed	257.8889	30.9450	256.7960	To	258.9817*
trimmed	250.0570	29.8788	248.6100	To	251.5039

Table A-9
Trimmed and Untrimmed Mean scores for the
1990 NAEP Reading Assessment

		Mean	Standard Deviation	95% Confidence Interval	
TOTAL					
	untrimmed	255.1148	36.6071	254.3459 To	255.8838*
	trimmed	252.3169	36.3202	251.5113 To	253.1226
RACE					
WHITE-	untrimmed	261.4020	35.8306	260.4909 To	262.3132*
	trimmed	258.7567	35.8484	257.7898 To	259.7237
BLACK-	untrimmed	240.4781	32.6796	238.7594 To	242.1969
	trimmed	237.9069	31.5594	236.1757 To	239.6382
HISPANIC-	untrimmed	236.1319	34.2240	234.0111 To	238.2528
	trimmed	233.8969	33.4040	231.7403 To	236.0535
ASIAN-	untrimmed	271.1322	34.3519	266.8611 To	275.8033
	trimmed	269.3603	34.9859	264.2224 To	274.4982
Gender					
MALE-	untrimmed	247.4656	37.0184	246.3852 To	248.5460*
	trimmed	244.9866	36.5872	243.8683 To	246.1048
FEMALE-	untrimmed	263.3403	34.3108	262.3019 To	264.3787*
	trimmed	260.4813	34.2241	259.3774 To	261.5853
REGION					
NORTHEAST-	untrimmed	261.8622	37.4867	260.0960 To	263.6284
	trimmed	259.2994	37.5020	257.4209 To	261.1779
SOUTHEAST-	untrimmed	251.1049	35.1739	249.6142 To	252.5956
	trimmed	248.4675	34.7004	246.9240 To	250.0110
CENTRAL-	untrimmed	255.1708	36.2098	253.6560 To	256.6857
	trimmed	252.9316	36.2544	251.3418 To	254.5215
WEST-	untrimmed	253.8901	36.9042	252.4810 To	255.2992*
	trimmed	250.4070	36.2641	248.9343 To	251.8798

		Standard			
	Mean	Deviation	95% Confidence Interval		
Type of Community					
EXTREME RURAL- untrimme	252.3159	35.8644	250.1345	To	254.4974
trimmed	249.9025	35.4054	247.6262	To	252.1789
DISADVANTAGED URBAN- un	246.0557	35.8392	243.6655	To	248.4460
trimmed	243.9775	35.6240	241.5011	To	246.4539
ADVANTAGED URBAN- untri	267.7908	36.7497	265.4672	To	270.1143
trimmed	264.6488	37.0364	262.1075	To	267.1902
OTHER (NON-EXTREME)- un	254.8659	36.2867	253.9350	To	255.7969*
trimmed	252.0904	35.9952	251.1176	To	253.0632
Size of Community					
EXTREME RURAL- untrimme	252.3159	35.8644	250.1345	To	254.4974
trimmed	249.9025	35.4054	247.6262	To	252.1789
LOW METROPOLITAN- untri	246.0557	35.8392	243.6655	To	248.4460
trimmed	243.9775	35.6240	241.5011	To	246.4539
HIGH METROPOLITAN- untr	267.7908	36.7497	265.4672	To	270.1143
trimmed	264.6488	37.0364	262.1075	To	267.1902
MAIN BIG CITY- uncrimme	253.3323	36.7914	250.7013	To	255.9634
trimmed	249.3146	36.4602	246.5319	To	252.0972
URBAN FRINGE- untrimmed	257.1468	35.3301	254.8339	To	259.4597
trimmed	254.9849	35.1295	252.5701	To	257.3997
MEDIUM CITY- untrimmed	252.7267	37.0709	250.4671	To	254.9864
trimmed	249.2954	36.6629	246.9307	To	251.6602
SMALL PLACE- untrimmed	255.2859	36.1394	254.0233	To	256.5484
trimmed	252.8128	35.8307	251.4987	To	254.1269

Table A-10
Trimmed and Untrimmed Mean scores for the
1992 NAEP Estimation Assessment

		Mean	Standard Deviation	95% Confidence Interval	
TOTAL					
	untrimmed	265.4028	27.3198	264.3128	To 266.4927
	trimmed	265.0662	24.0090	263.9798	To 266.1527
RACE					
WHITE-	untrimmed	273.0980	24.1738	271.9344	To 274.2615
	trimmed	271.0853	21.4844	269.9229	To 272.2477
BLACK-	untrimmed	242.0890	23.7636	239.7260	To 244.4521
	trimmed	245.1699	21.3878	242.7047	To 247.6351
HISPANIC-	untrimmed	249.2516	24.1316	246.3266	To 252.1767
	trimmed	251.8022	22.0724	248.7222	To 254.8821
ASIAN-	untrimmed	277.8979	23.7417	272.3442	To 283.4516
	trimmed	277.9566	22.0261	271.9057	To 284.0075
Gender					
MALE-	untrimmed	268.4500	26.4312	267.0359	To 269.8640
	trimmed	267.3207	23.4104	265.9085	To 268.7328
FEMALE-	untrimmed	261.5785	27.9414	259.9035	To 263.2535
	trimmed	262.1581	24.4689	260.4811	To 263.8350
REGION					
NORTHEAST-	untrimmed	264.0344	33.3317	261.2203	To 266.8484
	trimmed	265.7113	27.9296	262.9697	To 268.4528
SOUTHEAST-	untrimmed	258.3622	25.1983	256.2928	To 260.4315
	trimmed	258.3844	22.7220	256.2655	To 260.5033
CENTRAL-	untrimmed	271.1284	24.0727	269.2862	To 272.9706
	trimmed	269.6818	22.1362	267.7899	To 271.5738
WEST-	untrimmed	266.9531	25.1270	265.0091	To 268.8971
	trimmed	265.6031	22.4754	263.6435	To 267.5627

		Standard			
	Mean	Deviation	95% Confidence Interval		
Type of Community					
EXTREME RURAL- untrimme	269.6590	23.2296	266.6392	To	272.6787
trimmed	267.9534	21.3151	264.8682	To	271.0387
DISADVANTAGED URBAN- un	241.7129	26.6084	238.6923	To	244.7336
trimmed	246.9463	23.5188	243.7458	To	250.1468
ADVANTAGED URBAN- untri	283.5862	22.2846	281.3379	To	285.8345
trimmed	280.2231	19.8940	277.8861	To	282.5601
OTHER (NON-EXTREME)- un	264.8944	25.5197	263.6044	To	266.1843
trimmed	264.2431	23.0161	262.9406	To	265.5456
Size of Community					
EXTREME RURAL- untrimme	269.6590	23.2296	266.6392	To	272.6787
trimmed	267.9534	21.3151	264.8682	To	271.0387
LOW METROPOLITAN- untri	241.7129	26.6084	238.6923	To	244.7336
trimmed	246.9463	23.5188	243.7458	To	250.1468
HIGH METROPOLITAN- untr	283.5862	22.2846	281.3379	To	285.8345
trimmed	280.2231	19.8940	277.8861	To	282.5601
MAIN BIG CITY- untrimme	259.8210	25.1505	255.8922	To	263.7498
trimmed	260.1624	23.3422	255.7711	To	264.5537
URBAN FRINGE- untrimmed	267.0984	24.6783	264.1565	To	270.0404
trimmed	264.6783	21.0911	261.8175	To	267.5392
MEDIUM CITY- untrimmed	260.3434	25.6786	257.9637	To	262.7230
trimmed	259.9240	22.9687	257.5392	To	262.3089
SMALL PLACE- untrimmed	268.5132	25.1978	266.5316	To	270.4948
trimmed	267.9096	23.1399	265.9181	To	269.9011

Table A-11
Trimmed and Untrimmed Mean scores for the
1992 NAEP Mathematics Assessment

		Mean	Standard Deviation	95% Confidence Interval	
TOTAL					
	untrimmed	261.7970	36.6165	261.0895 To	262.5045
	trimmed	261.5125	35.6519	260.6156 To	262.4095
RACE					
WHITE-	untrimmed	272.1966	32.4825	271.4416 To	272.9516
	trimmed	276.4293	29.9987	272.8605 To	279.9980
BLACK-	untrimmed	229.6070	30.0431	228.1644 To	231.0496
	trimmed	230.6103	30.4438	228.7262 To	232.4944
HISPANIC-	untrimmed	240.4730	33.1076	238.4897 To	242.4562
	trimmed	242.6372	33.6445	240.0099 To	245.2645
ASIAN-	untrimmed	282.2322	38.3069	277.2553 To	287.2090
	trimmed	272.1503	37.0962	265.1380 To	279.1626
Gender					
MALE-	untrimmed	261.4486	37.0301	260.4648 To	262.4324
	trimmed	262.2672	36.2317	261.0169 To	263.5174
FEMALE-	untrimmed	262.1885	36.1460	261.1705 To	263.2065
	trimmed	260.6555	34.9686	259.3696 To	261.9415
REGION					
NORTHEAST-	untrimmed	261.1334	38.0964	259.5156 To	262.7512
	trimmed	260.7219	36.9444	258.6643 To	262.7794
SOUTHEAST-	untrimmed	253.8782	36.0573	252.4872 To	255.2693
	trimmed	254.4826	35.0503	252.7433 To	256.2218
CENTRAL-	untrimmed	268.8319	34.7253	267.5181 To	270.1458
	trimmed	268.7051	33.9432	267.0168 To	270.3933
WEST-	untrimmed	262.8282	36.3157	261.5034 To	264.1530
	trimmed	261.9672	35.4871	260.2854 To	263.6490

		Standard			
	Mean	Deviation	95% Confidence Interval		
Type of Community					
EXTREME RURAL- untrimme	262.8604	31.7564	260.8607	To	264.8602
trimmed	263.6050	30.3965	261.1629	To	266.0471
DISADVANTAGED URBAN- un	232.8431	32.8711	230.7207	To	234.9656
trimmed	234.6932	32.4309	232.0213	To	237.3651
ADVANTAGED URBAN- untri	286.6685	33.1803	284.5597	To	288.7773
trimmed	286.6764	30.0749	284.1317	To	289.2212
OTHER (NON-EXTREME)- un	262.0661	35.5535	261.2582	To	262.8740
trimmed	261.6110	34.9372	260.5745	To	262.6475
Size of Community					
EXTREME RURAL- untrimme	262.8604	31.7564	260.8607	To	264.8602
trimmed	263.6050	30.3965	261.1629	To	266.0471
LOW METROPOLITAN- untri	232.8431	32.8711	230.7207	To	234.9656
trimmed	234.6932	32.4309	232.0213	To	237.3651
HIGH METROPOLITAN- untr	286.6685	33.1803	284.5597	To	288.7773
trimmed	286.6764	30.0749	284.1317	To	289.2212
MAIN BIG CITY- untrimme	254.8750	37.2305	252.4978	To	257.2522
trimmed	255.2426	37.0525	252.2042	To	258.2809
URBAN FRINGE- untrimmed	266.8053	36.5187	264.8599	To	268.7507
trimmed	264.8824	35.3132	262.3574	To	267.4074
MEDIUM CITY- untrimmed	260.4308	36.5363	258.7218	To	262.1398
trimmed	259.3860	36.6548	257.1354	To	261.6366
SMALL PLACE- untrimmed	263.0241	33.7743	261.8857	To	264.1625
trimmed	263.3259	32.9563	261.8871	To	264.7647

Table A-12
Trimmed and Untrimmed Mean scores for the
1992 NAEP Reading Assessment

		Mean	Standard Deviation	95% Confidence Interval	
TOTAL					
	untrimmed	254.5656	34.8593	253.9576	To 255.1736*
	trimmed	249.7907	36.1853	249.0254	To 250.5559
RACE					
WHITE-	untrimmed	262.9360	31.6278	262.2722	To 263.5999*
	trimmed	259.4507	32.7327	258.6094	To 260.2920
BLACK-	untrimmed	230.4825	31.6687	229.1100	To 231.8550*
	trimmed	224.7250	32.0534	223.0962	To 226.3538
HISPANIC-	untrimmed	234.8976	34.4979	233.0343	To 236.7610*
	trimmed	229.1909	35.1691	226.9518	To 231.4301
ASIAN-	untrimmed	266.1564	35.8572	262.3000	To 270.0127
	trimmed	259.2922	35.7417	254.2553	To 264.3291
Gender					
MALE-	untrimmed	248.3557	34.4637	247.5251	To 249.1864*
	trimmed	244.0236	35.7578	243.0033	To 245.0438
FEMALE-	untrimmed	261.3947	34.0069	260.5351	To 262.2542*
	trimmed	256.8260	35.4562	255.7085	To 257.9434
REGION					
NORTHEAST-	untrimmed	257.4283	35.2026	256.0859	To 258.7707*
	trimmed	252.6433	36.6077	250.9522	To 254.3344
SOUTHEAST-	untrimmed	247.3552	34.8481	246.1278	To 248.5827*
	trimmed	242.6807	36.2815	241.1522	To 244.2092
CENTRAL-	untrimmed	259.6906	32.8467	258.5693	To 260.8119*
	trimmed	256.1321	33.6266	254.7519	To 257.5124
WEST-	untrimmed	253.9712	35.3404	252.8145	To 255.1279*
	trimmed	247.9908	36.8726	246.4958	To 249.4858

		Standard		
	Mean	Deviation	95% Confidence Interval	
Type of Community				
EXTREME RURAL- untrimme	257.8158	31.4822	255.8452	To 259.7864
trimmed	254.9434	32.8105	252.4774	To 257.4095
DISADVANTAGED URBAN- un	230.8832	33.2964	229.0792	To 232.6873*
trimmed	225.4694	33.8869	223.3238	To 227.6150
ADVANTAGED URBAN- untri	277.1938	30.3505	275.4536	To 278.9339*
trimmed	272.9212	31.6332	270.5964	To 275.2461
OTHER (NON-EXTREME)- un	254.7134	33.8456	254.0204	To 255.4064*
trimmed	250.3265	35.2128	249.4522	To 251.2007
Size of Community				
EXTREME RURAL- untrimme	257.8158	31.4822	255.8452	To 259.7864
trimmed	254.9434	32.8105	252.4774	To 257.4095
LOW METROPOLITAN- untri	230.8832	33.2964	229.0792	To 232.6873*
trimmed	225.4694	33.8869	223.3238	To 227.6150
HIGH METROPOLITAN- untr	277.1938	30.3505	275.4536	To 278.9339*
trimmed	272.9212	31.6332	270.5964	To 275.2461
MAIN BIG CITY- untrimme	247.5926	34.6502	245.6001	To 249.5851*
trimmed	242.4386	35.8447	239.9901	To 244.8872
URBAN FRINGE- untrimmed	259.1682	33.5117	257.5802	To 260.7563*
trimmed	253.5277	34.2656	251.5227	To 255.5327
MEDIUM CITY- untrimmed	252.1950	34.4519	250.7723	To 253.6176*
trimmed	248.0176	35.9363	246.2366	To 249.7986
SMALL PLACE- untrimmed	252.7290	34.5433	251.4298	To 254.0283
trimmed	256.2848	32.9817	255.2666	To 257.3031

Appendix 2: Software Routines

Software Kernels Used in the Analysis

QuickBasic programs were used to reorganize the data and to insert fit statistics into the NAEP datafile. The actual analysis was conducted using the SPSS control cards supplied by ETS, modified to reflect the inclusion of fit statistics. The analysis of the trial state assessment also involved the recomputation of the weights as subsets of the data were used.

The three basic kernels are:

RAW2IRT - reads in NAEPLAY.TXT outputs NAEP2.IRT which is the parameters for the items of interest. This gets broken into test.IRT files.
TXT2IRT2 - reads in test.IRT files and test.SPX file, generates a test2.IRT file which has the irt parameters and the item location.
INSERT - reads in the test2.IRT and test.DAT files, calculates fit and creates a new datafile with the fit statistic inserted.

The programs take advantage of the consistency in the NAEP dataset layout.

'RAW2IRT.BAS

```
OPEN "I", 1, "NAEPLAY.TXT"
OPEN "O", 2, "NAEP2.irt"
WHILE NOT EOF(1)
  LINE INPUT #1, A$
  IF INSTR(A$, "ATTACHMENT:") THEN
    PRINT A$
    PRINT #2, A$
  END IF
  IF MID$(A$, 186, 1) <> " " THEN
    IF INSTR(A$, "(RATER 1)") THEN B$ = "1" ELSE B$ = " "
    PRINT MID$(A$, 1, 7); MID$(A$, 94, 10); MID$(A$, 114, 10);
    PRINT MID$(A$, 154, 10); " "; MID$(A$, 178, 1); MID$(A$, 186, 3); B$
    PRINT #2, MID$(A$, 1, 7); MID$(A$, 94, 10); MID$(A$, 114, 10);
    PRINT #2, MID$(A$, 154, 10); " "; MID$(A$, 178, 1); MID$(A$, 186, 3); B$
  END IF
WEND
CLOSE
```

'TXT2IRT2.BAS

```
INPUT "which", wh$
wh$ = RTRIM$(LTRIM$(wh$))
OPEN "i", 1, wh$ + ".IRT"
OPEN "i", 2, wh$ + "13.SPX"
OPEN "O", 3, wh$ + "2.IRT"
REDIM c$(1050)
CLS
```

```

NIJ = 0: OK = 1
WHILE NOT EOF(2)
  LINE INPUT #2, cc$
  cc$ = LTRIM$(RTRIM$(cc$))
  IF LEFT$(cc$, 9) = "VALUE LAB" THEN OK = 0
  IF LEFT$(cc$, 6) = "RECODE" THEN OK = 1
  IF OK = 1 THEN
    NIJ = NIJ + 1
    c$(NIJ) = cc$
  END IF
WEND
CLOSE #2

WHILE NOT EOF(1)
  LINE INPUT #1, a$
  IT$ = LEFT$(a$, 7)
  NI = 0
  LOCA$ = ""
  R1$ = " "
  R2$ = " "
  R3$ = " "
  L$ = " "
  FOR i = 1 TO NIJ
    IF INSTR(c$(i), IT$) THEN
      NI = NI + 1
      IF NI = 1 THEN
        LOCA$ = MID$(c$(i), INSTR(c$(i), IT$) + 9, 10)
        LOCA$ = STR$(VAL(LOCA$))
        L$ = " "
        RSET L$ = LOCA$
      END IF
      IF NI = 2 THEN
        IF INSTR(c$(i), "(RATER 1)") THEN R1$ = "1" ELSE R1$ = " "
        IF INSTR(c$(i), "(CALC USE)") THEN R2$ = "1" ELSE R2$ = " "
      END IF
      IF INSTR(c$(i), "RECODE") THEN
        key$ = " "
        jj = INSTR(c$(i), "=1")
        IF jj THEN
          key$ = MID$(c$(i), jj - 1, 1)
          IF MID$(c$(i), jj - 7, 4) = "THRU" THEN
            key$ = MID$(c$(i), jj - 9, 1) + key$
          ELSE
            key$ = key$ + " "
          END IF
        END IF
      END IF
      IF key$ = " " THEN 'NEXT LINE
        jj = INSTR(c$(i + 1), "=1")
        IF jj THEN
          key$ = MID$(c$(i + 1), jj - 1, 1)
          IF MID$(c$(i + 1), jj - 7, 4) = "THRU" THEN
            key$ = MID$(c$(i + 1), jj - 9, 1) + key$
          ELSE
            key$ = key$ + " "
          END IF
        END IF
      END IF
    END IF
  NEXT i
  EXIT FOR

```

```
        END IF
    END IF
NEXT i
IF NI > 0 THEN
    'PRINT a$
    c = VAL(MID$(a$, 28, 10))
    IF c = 0 THEN r3$ = "1" ELSE r3$ = " "
    bb$ = MID$(a$, 40, 3)
    MID$(a$, 39, 5) = key$ + bb$
    PRINT LEFT$(a$, 44) + L$ + R1$ + R2$ + r3$
    PRINT #3, LEFT$(a$, 44) + L$ + R1$ + R2$ + r3$
    ' r1 rater
    ' r2 calc
    ' r3 constructed response
END IF
WEND
```

```
'INSERT.BAS
' INSERT fit in new data file
```

```
DECLARE FUNCTION fixfit$ (fit!)
DEFINT I-N
```

```
INPUT "which?:"; wh$
OPEN "i", 1, wh$ + ".DAT"
OPEN "O", 3, wh$ + "2.DAT"
OPEN "i", 2, wh$ + "2.IRT"
INPUT "theta start:"; ithetast
INPUT "N scales"; nscales
icol = 200 'where to insert fit stats

'ithetast = 696 ' theta start column (y21 reading)
ithsz = 6 ' theta size
```

```
REDIM a(250)
REDIM B(250)
REDIM C(250)
REDIM ICOR1$(250), ILOCA(250), irat$(250), icalc$(250), icor2$(250)
REDIM scale(250), icons$(250)
REDIM atheta(5), afit(7), aTOP(7), aBOT(7)
REDIM asum(7), ant(7)
ni = 0
```

```
' gather item info
WHILE NOT EOF(2)
    LINE INPUT #2, a$
    ni = ni + 1
    a(ni) = VAL(MID$(a$, 8, 10))
    B(ni) = VAL(MID$(a$, 18, 10))
    C(ni) = VAL(MID$(a$, 28, 10))
    ICOR1$(ni) = MID$(a$, 39, 1)
    icor2$(ni) = MID$(a$, 40, 1)
    scale(ni) = VAL(MID$(a$, 41, 1))
    ILOCA(ni) = VAL(MID$(a$, 44, 5))
    irat$(ni) = MID$(a$, 49, 1)
    icalc$(ni) = MID$(a$, 50, 1)
    icons$(ni) = MID$(a$, 51, 1)
```

```

WEND
CLOSE #2
nfit = 0

' for each record in the data file
WHILE NOT EOF(1) 'AND NFIT < 100
  th = 0
  TOP = 0
  BOT = 0
  FOR j = 1 TO 7
    aTOP(j) = 0
    aBOT(j) = 0
    afit(j) = 0
  NEXT j
  LINE INPUT #1, a$
  FOR k = 1 TO nscales
    NTH = 0
    th = 0
    ' calculate theta
    FOR j = 1 TO 5
      IF (MID$(a$, ithetast - ithsz + j * ithsz, ithsz)) <> " " THEN
        NTH = NTH + 1
        th = th + VAL(MID$(a$, ithetast - ithsz + j * ithsz, ithsz)) / 10000
      END IF
    NEXT j

    IF NTH > 0 THEN theta = th / NTH ELSE theta = -9.999
    IF theta = -9.999 THEN GOTO drop
    'PRINT th, theta; a
    atheta(k) = theta
  NEXT k

  ' for each item, calculate p, u and then fit
  FOR i = 1 TO ni
    RES$ = MID$(a$, ILOCA(i), 1)
    isc = scale(i)
    IF VAL(RES$) <> 0 THEN
      U = 0
      IF ((RES$ = "1") AND (ICOR1$(i) = "0"))
        OR (RES$ = ICOR1$(i)) OR RES$ = icor2$(i) THEN U = 1
      P = C(i) + (1 - C(i)) / (1 + EXP(-1.7 * a(i) * (atheta(isc) - B(i))))
      aTOP(1) = aTOP(1) + (U - P) ^ 2
      aBOT(1) = aBOT(1) + P * (1 - P)
      IF irat$(i) = "1" THEN
        aTOP(2) = aTOP(2) + (U - P) ^ 2
        aBOT(2) = aBOT(2) + P * (1 - P)
      ELSE
        aTOP(3) = aTOP(3) + (U - P) ^ 2
        aBOT(3) = aBOT(3) + P * (1 - P)
      END IF
      IF icalc$(i) = "1" THEN
        aTOP(4) = aTOP(4) + (U - P) ^ 2
        aBOT(4) = aBOT(4) + P * (1 - P)
      ELSE
        aTOP(5) = aTOP(5) + (U - P) ^ 2
        aBOT(5) = aBOT(5) + P * (1 - P)
      END IF
      IF icons$(i) = "1" THEN

```

```

    aTOP(6) = aTOP(6) + (U - P) ^ 2
    aBOT(6) = aBOT(6) + P * (1 - P)
  ELSE
    aTOP(7) = aTOP(7) + (U - P) ^ 2
    aBOT(7) = aBOT(7) + P * (1 - P)
  END IF
  PRINT theta; B(i); " "; U; P, RES$; ICOR1$(i); icor2$(i)
END IF
NEXT i

' calculate fit
FOR j = 1 TO 7
  IF aBOT(j) > 0 THEN afit(j) = aTOP(j) / aBOT(j) ELSE afit(j) = -9.999
  IF afit(j) > 10 THEN afit(j) = 9.999
  IF afit(j) <> -9.999 THEN
    asum(j) = asum(j) + afit(j)
    ant(j) = ant(j) + 1
  END IF
NEXT j
SUMFIT = SUMFIT + afit(1)
nfit = nfit + 1
PRINT nfit

' fix at 3 decimals
FOR j = 1 TO 7
  fit$ = fixfit$(afit(j))
' insert in A$ and write
  MID$(a$, icol + 8 * (j - 1), 8) = fit$
NEXT j
PRINT #3, a$

drop:
IF INKEY$ = CHR$(27) THEN GOTO done
WEND

done:
PRINT "MEAN= "; SUMFIT / nfit
FOR i = 1 TO 7
  IF ant(i) > 0 THEN PRINT i, asum(i) / ant(i), ant(i) ELSE PRINT
NEXT i
CLOSE

FUNCTION fixfit$ (fit)
  a = (CLNG(fit * 1000)) / 1000
  F$ = STR$(a)
  IF INSTR(F$, ".") = 0 THEN F$ = F$ + "."
  WHILE (LEN(F$) - INSTR(F$, ".")) < 3
    F$ = F$ + "0"
  WEND
  fit$ = " "
  RSET fit$ = F$
  fixfit$ = fit$
END FUNCTION

```

References

- Alexander, L. & James, H.T. (1987). *The Nation's Report Card: Improving the assessment of student achievement*. Washington, DC: National Academy of Education.
- Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinees ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Bock, R.D. (1985) Designing the National Assessment of Educational Progress to Serve a Wider Community of Users: A Position Paper. Paper commissioned by the Study Group on the National Assessment of Student Achievement. ERIC Document Number ED279664.
- Chang, S.T. & W.L. Bashaw (1984) Characteristics of Anchor Tests. Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
- Cliff, N. (1979) Test Theory without True Scores? *Psychometrika*, 44(4), 373-393.
- Cronbach, L.J. (1972) Test validation. In R.L. Thorndike (ed) *Educational Measurement*, 2nd edition. Washington, DC: American Council on Education.
- Donlon, T.F., & Fischer, F.E. (1968) An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement*, 28, 105-113.
- Doss, D.A. (1981) Will removing a few bad apples save the barrel? ERIC Document Number ED202916.
- Doss, D.A. & G. Ligon (1985) Empty Bubbles: What Test Form Did They Take? Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
- Drasgow, F. & M.V. Levine (1986) Optimal detection of certain forms of inappropriate test scores *Applied Psychological Measurement*, 10, 59-67.
- Drasgow, F., et.al (1984) Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices. ERIC Document Number ED246072.
- Drasgow, F., M.V. Levine, & M.E. McLaughlin (1987) Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 1, 59-79.
- Drasgow, F., M.V. Levine, & M.E. McLaughlin (1987) Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 1, 59-79.
- Educational Testing Service (May, 1991). 1990 trial state assessment secondary-use data files user guide. Princeton, NJ: author.

- Frary, R.B. (1980) The effect of misinformation, partial information, and guessing on exoected multiple choice test item scores, *Applied Psycho'ogical Measurement*, 4(1), 79-90.
- Frary, R.B. (1982) A Comparison among person-fit measures. Paper presented at the annual meeting of the American educational Research Association, New York, March 19-23.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. *American Psychologist*, 18, 519-521.
- Gustafsson, J.E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33, 205-233.
- Haertel, E.H., et al. (1989). *Report of the NAEP technical review panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level NAEP comparisons*. Washington, DC: National Center for Education Statistics.
- Harnisch, D.L., & Linn, R.L. (1981) Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 3, 133-146.
- Hattie, J. & H.J. Rogers (1986) Factor Models for Assessing the Relationship between Creativity and Intelligence. *Journal of Educational Psychology*, 78(6), 482-485.
- Hodgkinson, H. (1991) Reform versus Reality. *Phi Delta Kappan*, 73(1), 8-16.
- Johnson, E. G. & Allen, N.L. (1991). The NAEP 1990 Technical Report. Washington, DC: National Center for Education Statistics.
- Klauer, K.C. (1991) Exact and Best Confidence Intervals for the Ability Parameter of the Rasch Model. *Psychometrika*, 56(3), 535-547.
- Levine, M.V. & D. Rubin (1979) Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Levine, M.V. & F. Drasgow (1982) Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical Statistical Psychology*, 35, 42-56.
- Levine, M.V. & F. Drasgow (1988) Optimal Appropriateness Measurement. *Psychometrika*, 53(2), 161-176.
- Levine, M.V. & F. Drasgow (1987) The Relation between Incorrect Option Choice and Estimated Ability. *Educational and Psychological Measurement*, 43(3), 675-685.
- Linquist, E.F. (1951) *Design and analysis of experiments in education and psychology*. Boston: Houghton Mifflin.
- Lord, F.M. & M.R. Novick (Eds., 1968), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McKinley, R.L. & C.N. Mills (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9, 49-57.
- Mead, R.J. (1976) *Analysis of fit to the Rasch model*. Unpublished doctoral dissertation, University of Chicago.
- Mead, R.J. & Kreines, D.C. (1980) *Person fit analysis with the dichotomous Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April.

- Messick, S., Beaton, A., & F.M. Lord (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era.* (NAEP Report 83-1) Princeton, NJ: Educational Testing Service.
- Mislevy, R.J. (1988). *Final report: Exploiting collateral information in the estimation of item parameters.* (ETS Research Report, RR-88-53-ONR). Princeton, NJ: Educational Testing Service.
- Molenaar, I.W. & H. Hoijtink (1990) The Many Null Distributions of Person Fit Indices. *Psychometrika*, 55(1), 75-106
- Mullis, I.V.S., Dossey, J.A., Owen, E.H., & G.W. Phillips (June, 1991). The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. Washington, DC: National Center for Education Statistics.
- Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests.* Kobenhaven: Denmark's Paedagogiske Institut. (Reprinted by University of Chicago Press, 1980.)
- Reise, S.P. (1990) A comparison of Item- and Person-Fit Methods of Assessing Model Data Fit in IRT, *Applied Psychological Measurement*, 14, 2, 127-137.
- Reise, S.P. & A.M. Due (1991) The Influence of Test Characteristics on the Detection of Aberrant Response Patterns. *Applied Psychological Measurement*, 15(3), 217-226.
- Rudner, L.M. (1983a) Individual Assessment Accuracy. *Journal of Educational Measurement*, 20(3), 207-219.
- Rudner, L.M. (1983b) A Closer Look at Latent Trait Parameter Invariance. *Educational and Psychological Measurement*, 43(4), 951-955.
- Rudner, L.M., Getson, P. & Knight, D. (1980). Item Bias Detection Techniques, *Journal of Educational Statistics*, 5, 213-233.
- Sato, T. (1975) [The construction and interpretation of S-P Tables.] Tokyo: Meiji Tosho, (in Japanese; cited in Harnisch and Linn, 1981).
- Schmitt, A.P. & L. Crocker (1984) The Relationship between Test Anxiety and Person Fit Measures. Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984)
- Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, 359-372.
- Smith, R.M. (1991) The Distributional Properties of the Rasch Item Fit Statistics, *Educational and Psychological Measurement*, 51, 541-565.
- Tatsuoka, K. & M. Birenbaum (1979) The Danger of Relying Solely on Diagnostic Adaptive Testing When Prior and Subsequent Instructional Methods Are Different. ERIC Document number ED183608.
- Tatsuoka, K.K. (1984) Caution Indices Based on Item Response Theory. *Psychometrika* 49(1), 95-110.
- Tatsuoka, K.K., & Linn, R.L. (1983) Indices for detecting unusual response patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, 7, 1, 81-96.

- Tatsuoka, K.K., & Tatsuoka, M.M. (1982) Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, 7(3), 215-231.
- Trabin, T.E., & Weiss, D.J. (1979) The person response curve: Fit of individuals to item characteristic curve models (RR 79-7). Minneapolis, Minn.: University of Minnesota, Department of Psychology, Psychometric Methods Program, December.
- van der Flier, H. (1977) Environmental factors and deviant response patterns. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets and Zeitlinger, B.V.
- Wainer, H., Morgan, A., & Gustafsson, J.E. (1980). A review of estimation procedures for the Rasch model with a view towards longish tests. *Journal of Educational Statistics*, 1, 35-69.
- Westfall, P.J-L. & A.G. D'Costa (1987) Improving Prediction by Correcting Test Scores for Person Disturbances Using the Rasch Model. ERIC Document Number ED319768.
- Wiggins, G. (1990) The case for authentic assessment. *ERIC/TM Digest*, TM-90-5.
- Wright, B. & M. Stone (1979) *Best Test Design*. Chicago: MESA Press.
- Wright, B.D. (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-115.
- Wright, B.D., & Panchapakesan, N.A. (1969) A procedure for sample free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B.D., Mead, R., & Bell, S. (1979) BICAL: Calibrating items with the Rasch model (RM-23b). University of Chicago.
- Yoes, M.E. & K.T. Ho (1991) The Degree of Person Misfit on a Nationally Standardized Achievement Test. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).

United States
Department of Education
Washington, DC 20208-1527

Official Business
Penalty for Private Use, \$300

Postage and Fees Paid
U.S. Department of Education
Permit No. G-17

Third Class



BEST COPY AVAILABLE